Cross-modal Representation Learning and Relation Reasoning for Bidirectional Adaptive Manipulation

Lei Li[1,3] , **Kai Fan**[2] and **Chun Yuan**[*3]

[1]Department of Computer Science and Technology, Tsinghua University
[2]Alibaba DAMO Academy, Alibaba Group Inc.
[3]Tsinghua Shenzhen International Graduate School, Peng Cheng Lab

Abstract

Since single-modal controllable manipulation typically requires supervision of information from other modalities or cooperation with complex software and experts, this paper addresses the problem of cross-modal adaptive manipulation (CAM). The novel task performs cross-modal semantic alignment from mutual supervision and implements bidirectional exchange of attributes, relations, or objects in parallel, benefiting both modalities while significantly reducing manual effort. We introduce a robust solution for CAM, which includes two essential modules, namely Heterogeneous Representation Learning (HRL) and Cross-modal Relation Reasoning (CRR). The former is designed to perform representation learning for cross-modal semantic alignment on heterogeneous graph nodes. The latter is adopted to identify and exchange the focused attributes, relations, or objects in both modalities. Our method produces pleasing cross-modal outputs on CUB and Visual Genome.

1 Introduction

Generating natural and meaningful outputs from one modality that semantically matches given supervision information from the other modalities is a challenging problem with vast potential applications, including image editing, language style transfer, and computer-aided design. Considering that a cross-modal model can receive inputs from multiple modalities simultaneously, it is a natural idea that the outputs of these models should also cover all the input modalities simultaneously. However, the current research focuses on the output of a specific modality [Nam *et al.*, 2018; Li *et al.*, 2019; Dhamo *et al.*, 2020]. For example, text-guided image manipulation generates natural-looking images from language descriptions [Li *et al.*, 2020a; Li *et al.*, 2020b], and image-guided text manipulation produces language conforming to grammar rules [Cornia *et al.*, 2019] from a conditional image. However, cross-modal adaptive manipulation for bidirectional generation on both modalities has received less attention.

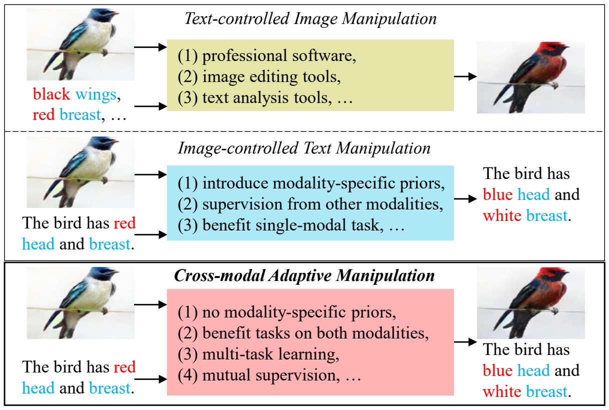

Text-controlled Image Manipulation
black wings,
red breast, …
(1) professional software,
(2) image editing tools,
(3) text analysis tools, …
Image-controlled Text Manipulation
The bird has red
head and breast.
(1) introduce modality-specific priors,
(2) supervision from other modalities,
(3) benefit single-modal task, …
The bird has
blue head and
white breast.
Cross-modal Adaptive Manipulation
The bird has red
head and breast.
(1) no modality-specific priors,
(2) benefit tasks on both modalities,
(3) multi-task learning,
(4) mutual supervision, …
The bird has
blue head and
white breast.

Figure 1: Highlights of the novel CAM task.

While one can introduce additional supervision to manipulate images or texts, it is challenging for a non-expert user to design appropriate rules, specific priors or adopt professional tools for single-modal manipulation. Additionally, the generation from single-modal manipulation has been mostly restricted to object-centric scenarios [Dhamo *et al.*, 2020] and the limited functionality of fixed or mobile devices. By comparison, cross-modal adaptive manipulation aims to achieve cross-modal mutual supervision with no need for specific priors, adaptively exchanging attributes, relations, or objects in input modalities, and generating outputs for both inputs by multi-task learning. Therefore, CAM can better fit cross-modal scenarios and inspire related research.

We specify the introduced task, i.e., consider the two critical modalities of vision and language. Given an image and a sentence, on the one hand, it allows modifying visual presentations (e.g., texture, color, and category) of the image according to the input sentence; on the other hand, it supports adjusting textual contexts (e.g., noun, adjective, and verb) of the sentence by the input image. The differences between our task and previous works are listed in Figure 1.

Our contributions are summarized as follows: **(1)** The weaknesses of single-modal controllable manipulation are addressed, and a novel task, namely cross-modal adaptive manipulation (CAM), is introduced, which has excellent potential for cross-modal applications. **(2)** A strong baseline is proposed to solve the CAM task. The HRL module performs

[*] Corresponding author

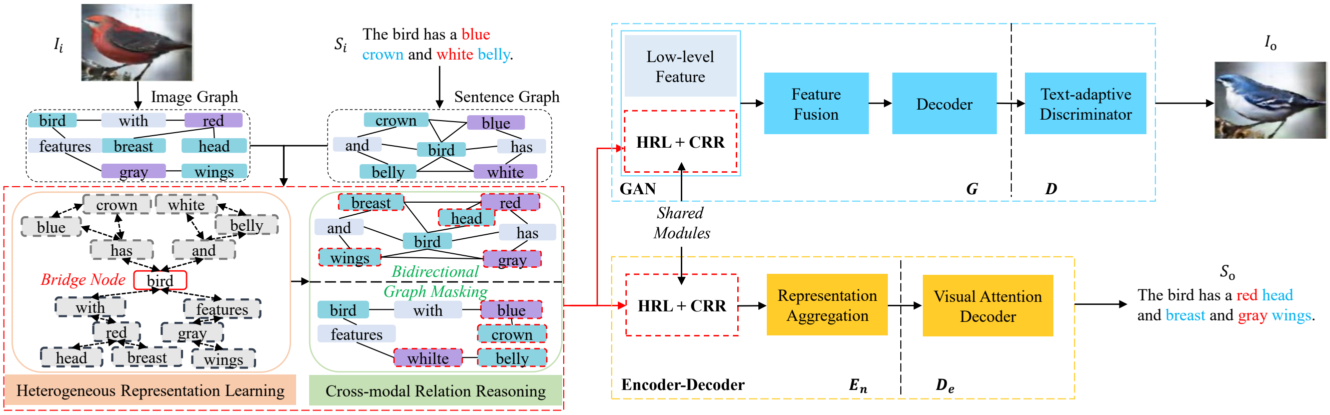

I_i
Image Graph
S_i
The bird has a blue crown and white belly.
Sentence Graph
bird
with
red
features
breast
head
gray
wings
crown
blue
and
bird
has
belly
white
Bridge Node
Heterogeneous Representation Learning
Bidirectional Graph Masking
Cross-modal Relation Reasoning
Low-level Feature
HRL + CRR
Feature Fusion
Decoder
Text-adaptive Discriminator
GAN
G
D
I_o
Shared Modules
Representation Aggregation
Visual Attention Decoder
Encoder-Decoder
E_n
D_e
S_o
The bird has a red head and breast and gray wings.

Figure 2: The pipeline of our method. We propose a graph representation learning approach for the novel cross-modal manipulation task.

cross-modal representation learning for semantic alignment, and the CRR module implements cross-modal relation reasoning in input modalities. **(3)** Experiments on CUB and Visual Genome verify that our approach outperforms the leading methods of single-modal controllable manipulation.

2 Related Work

Single-modal Controllable Manipulation. Distinct from a portion of generative networks that are typically uncontrollable, another line of research can be summarized as single-modal controllable manipulation, which mainly includes text-guided image manipulation and image-guided text manipulation. For text-guided image manipulation, [Nam *et al.*, 2018] utilized a text-adaptive generative adversarial network to semantically manipulate images while preserving independent contents in the original image. [Li *et al.*, 2020b] proposed a word-level discriminator to facilitate training a lightweight generator for image manipulation using natural language descriptions. For image-guided text manipulation, which is different from image caption (no additional text input) and rarely paid attention to, many applications can benefit from this task, such as the children's early teaching robot dialogue system.

However, the existing works only perform single-modal controllable manipulation for visual or language generation, thus only obtaining biased models which can only be applied in restricted applications. In this work, we introduce the novel CAM task for multi-modal generations, simultaneously benefiting the modalities of both inputs.

Graph Representation Learning. The latest developments in graph representation learning have revolutionized various applications as they are well suited for realistic scenarios. Generally, the main ways to apply graph networks for non-structural scenarios include visual and language are: (1) incorporate structural information from other modalities to improve the performance of the current modality. For example, using information from knowledge graphs to guide zero-shot recognition classification [Kampffmeyer *et al.*, 2019], or performing more delicate relation exploration for a more interpretable reasoning process [Wang *et al.*, 2018]; (2) infer or assume the relational structure defined on graphs in a single modality. For instance, [Yao *et al.*, 2019] regarded the documents and words as nodes to construct the corpus graph and used the Text GCN to guide the representation learning of nodes. [Liu *et al.*, 2019] conducted evidence aggregating and reasoning based on a fully connected evidence graph.

However, these approaches implement graph representation learning by introducing external supervision or only performing single-modal relation reasoning. In contrast, we explore cross-modal graph representation learning and bidirectional relation reasoning in the way of mutual supervision, leading to a novel multi-task learning framework.

3 Proposed Method

Task Definition. We introduce the cross-modal adaptive manipulation task: given an image I_i and a sentence S_i as inputs, on the one hand, the job semantically manipulates I_i according to S_i so that the visual presentation of the manipulated output I_o match the linguistic descriptions of S_i while preserving S_i-independent information in I_i; on the other hand, it adaptively manipulates S_i according to I_i so that language presentation of the manipulated output S_o match the visual contents of I_i while keeping I_i-independent information in S_i. Notably, the above two procedures are performed simultaneously.

An overview of our solution is shown in Figure 2, which could be split into two parallel pipelines. For controllable image manipulation, the pipeline is based on GAN [Goodfellow *et al.*, 2020] framework, in which the generator G is trained to produce $I_o = G(I_i, S_i)$, and the discriminator D is adopted to ensure that realistic images are generated to semantically match the conditional text. The pipeline adopts the Encoder-Decoder framework for controllable text manipulation, including the encoder E_n to capture the distribution of manipulated sentences and the decoder D_e to output the desired words sequentially. In the following, we first introduce the main contributions of this work, i.e., the HRL module and CRR module, which are shared as the first part of both G and E_n. After that, we will briefly describe the other features of GAN and the Encoder-Decoder framework, which are mainly based on the existing works.

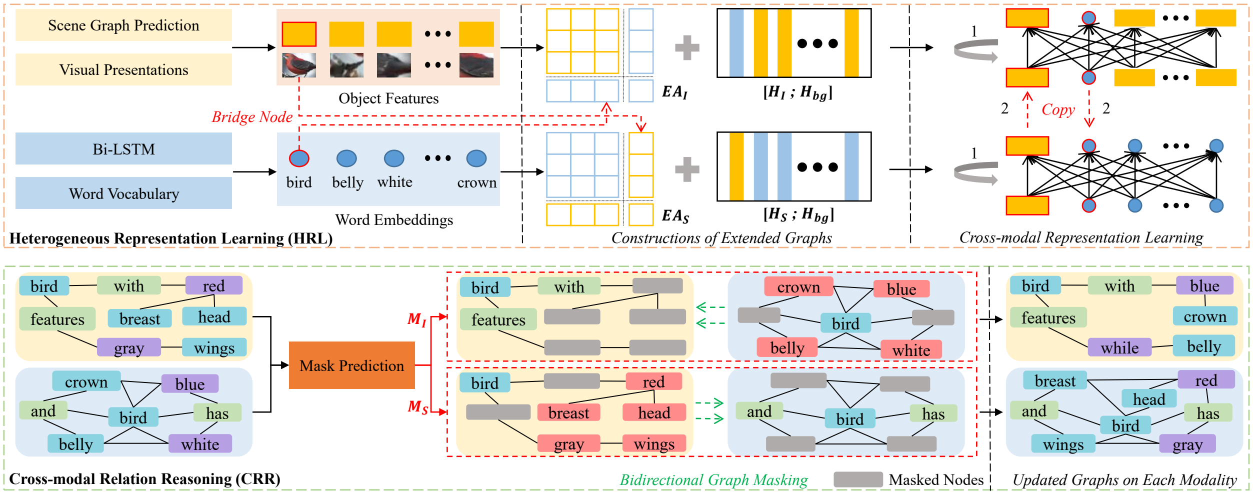


Figure 3: The two main modules of the proposed framework for multi-task learning.

3.1 Heterogeneous Representation Learning

Given both inputs, We first construct an image graph $\mathbf{G}_I$ for I_i and a sentence graph $\mathbf{G}_S$ for S_i, respectively. Specifically, we adopt a pre-trained model to discretize the input image to obtain multiple objects and treat each object as a visual node, eventually getting the $\mathbf{G}_I$ with its adjacency matrix A_I. During this period, we use a similar approach to build the $\mathbf{G}_S$ with its A_S.

We need to find bridges to link both input modalities to perform representation learning for cross-modal semantic alignment. Specifically, we search for objects co-occurrence in both modalities to serve as bridge nodes. As shown in Figure 3, the visual object '*bird*' and the word '*bird*' appear in the input modalities, respectively. In this case, we treat '*bird*' as a bridge object and create four kinds of nodes, including $bird_I$, $bird_S$, $bird_{IS}$ and $bird_{SI}$, which indicate the visual node of $\mathbf{G}_I$, the language node of $\mathbf{G}_S$, and the copies of $bird_I$ and $bird_S$, respectively. The first two nodes (denoted as reserved nodes) are kept in the built graphs of their original modalities, while the last two nodes will serve as bridge nodes to create heterogeneous graphs for cross-modal representation learning. Significantly, the node $bird_{IS}$ derived from visual modality is added to $\mathbf{G}_S$, while the node $bird_{SI}$ generated from language modality is incorporated into $\mathbf{G}_I$. We define $F(bird_{IS}) = F(bird_I), F(bird_{SI}) = F(bird_S)$ where $F(\cdot)$ returns the initial node features. A multi-layer perceptron (MLP) layer is then utilized to map the feature dimension of visual nodes to the same as the accordingly word embeddings. After that, we extend the adjacency matrix under each modality by introducing the bridge nodes from the other modality to obtain the extended version A_{EI} of the extended image graph $\mathbf{G}_{EI}$ and the corresponding A_{ES} of $\mathbf{G}_{ES}$. In this way, we obtain two heterogeneous graphs consisting of nodes from both input modalities.

Heterogeneous representation learning is then performed in the extended graphs of $\mathbf{G}_{EI}$ and $\mathbf{G}_{ES}$ in two steps. First, the representation aggregations based on regular graph convolutions [Kipf and Welling, 2016] is performed independently on each heterogeneous graph, i.e., this step in $\mathbf{G}_{EI}$ could be implemented by a layer-wise propagation rule:

$$\left(H_I^{l+1}; H_{bg}^{l+1}\right) = \text{ReLU}\left(\hat{A_{EI}}\left(H_I^l; H_{bg}^l\right) W^l\right),$$

where $\hat{A_{EI}}$ is the normalization of A_{EI}, W^l is a layer-specific trainable weight matrix, and $(;)$ means the row-wise concatenation. H_I^l indicates the feature matrix of visual nodes in the l-th layer, and H_{bg}^l is the feature vectors of introduced bridge nodes from the language modality. The same propagation process is performed in $\mathbf{G}_{ES}$.

After one-layer propagation of both graphs, we update the representation of all reserved nodes in both graphs periodically with the help of the copy mechanism. Take the reserved node $bird_I$ for example, in the second step, we merge the representation of related nodes through a linear transformation:

$$\bar{H}_{bird_I} = \lambda_1 H_{bird_I} + \lambda_2 \text{MLP}\left(H_{bird_{IS}} + H_{bird_S}\right),$$

where $H_{bird_{I(,IS,S)}}$ denotes the feature vectors of node $bird_{I(,IS,S)}$, and λ_1 and λ_2 are the weighting factors. After that, the representation of node $bird_I$ is updated by: $H_{bird_I} = \bar{H}_{bird_I}$. The above two-step learning process is iterated by L graph layers and finally, we obtain H_I^L and H_S^L. In this way, the introduced bridge nodes can learn the representation of both modalities and spread information in the heterogeneous graphs, progressively aligning the semantics of nodes from both input modalities while providing modal-consistent representation for subsequent relation reasoning.

3.2 Cross-modal Relation Reasoning

We introduce the bidirectional graph masking mechanism to implement cross-modal relation reasoning parallelly. Based on the updated representation of cross-modal nodes, we remove all the bridge nodes and predict two soft masks M_I and M_S for the original $\mathbf{G}_I$ and $\mathbf{G}_S$, which are formulated as:

$$h_{IS} = \text{MeanPooling}(H_I^L; H_S^L),$$

$$\hat{M}_{I(,S)} = H_{I(,S)}^L \cdot W_{I(,S)} \cdot h_{IS},$$

$$M_{I(,S)} = \text{Sigmoid}\left(\text{MaxPooling}\left(\hat{M}_{I(,S)}\right)\right),$$

Method	A	N	L_2	IS	FID	P-G	P-D	RPE (h)	IT (s)
SISGAN [Dong *et al.*, 2017]	2.34	2.42	0.43	2.41	16.81	-	-	-	-
AttnGAN [Xu *et al.*, 2018]	2.27	2.16	0.24	3.23	14.64	-	-	-	-
TaGAN [Nam *et al.*, 2018]	1.86	1.95	0.12	3.26	15.49	-	-	-	-
ManiGAN [Li *et al.*, 2020a]	1.76	1.63	0.12	8.16	10.53	41.1M	169.4M	0.12	2.45
Lightweight GAN [Li *et al.*, 2020b]	1.55	1.60	0.09	8.11	9.21	**18.5M**	71.8M	0.08	0.39
Ours	**1.33**	**1.49**	**0.07**	**8.34**	**8.74**	26.7M	**47.4M**	**0.07**	**0.35**

Table 1: Quantitative comparison for synthesized images on the CUB dataset. Accuracy (A) and Naturalness (N) are evaluated by users, and the values indicate the average ranking. For Accuracy, Naturalness, L_2, and FID, lower is better; for IS, higher is better. We also compare the number of parameters in generator (P-G) and discriminator (P-D), runtime per epoch (RPE), and inference time for generating 100 new modified images (IT). For P-G, P-D, RPE, and IT, lower is better. All methods are benchmarked on a single Nvidia GeForce RTX 3080 GPU.

Model	MAE ↓	SSIM ↑	LPIPS ↓	IS ↑	FID ↓
Cond-sg2im [Johnson *et al.*, 2018]	14.25	84.42	0.081	11.14±0.80	13.40
SIMSG (P) [Dhamo *et al.*, 2020]	13.82	83.98	0.077	10.61±0.37	16.69
SIMSG (GT) [Dhamo *et al.*, 2020]	8.53	87.57	0.051	12.07±0.97	7.54
Ours (P)	10.27	86.15	0.050	11.08±0.72	10.31
Ours (GT)	**7.41**	**89.02**	**0.047**	**12.39±0.75**	**7.06**

Table 2: Quantitative comparison for synthesized images on the Visual Genome. For MAE, LPIPS, and FID, lower is better; for SSIM and IS, higher is better. We report the results using ground truth scene graphs (GT) and predicted scene graphs (P).

where W_I, W_S are learnable matrices, and $(\cdot)$ indicates matrix multiplication. The process simultaneously performs global relations reasoning on both graphs. Each generated masking matrix enables the exchange of corresponding graph nodes in both modalities by masking some nodes in one modality and highlighting the corresponding (semantically similar) nodes in the other modality. For instance, consider the red node in the visual modality and the corresponding $blue$ node in the language modality, M_I tries to mask the red node and highlights the $blue$ node for generating the image with a '$blue$' object. While M_S attempts to do the opposite operation for generating the sentence with a 'red' word by masking the $blue$ node and highlighting the red node. By masking a portion of nodes and performing the representation aggregation of all highlighted nodes in both graphs, two advantages are obtained in this way: (1) the exchange of nodes is achieved across modalities without destroying the graph structure of each modality; (2) bidirectional relation reasoning is implemented through learning M_I and M_S in parallel on both modalities. Finally, the obtained soft masks will be multiplied by the node features and processed by a linear transformer to get the output:

$$H_{\mathbf{G}_{I(,S)}} = \text{MLP}\left((H_I; H_S) \otimes M_{I(,S)}\right),$$

where $\otimes$ denotes the element-wise product operation.

Other Parts for Manipulations. For text-guided image manipulation, GAN is adopted while its generator mainly contains the two modules mentioned above, one additional feature fusion part and a light decoder. We adopt the VGG-16 network to encode the input image, and the obtained features are combined with the representation of the node for fusion. The decoder, which consists of a few upsampling layers, is adopted for image generation. Additionally, we adopt the text-adaptive discriminator introduced in [Nam *et al.*, 2018] to force the generator to receive feedback from each local discriminator for each visual attribute.

We adopt the Encoder-Decoder framework for image-guided text manipulation, whose encoder contains the two modules mentioned earlier and a representation aggregation part and whose decoder is composed of a visual attention-based network. Specifically, we perform the representation aggregation process to obtain the attention-guided representation of each node in the language modality. The decoder generates one word at each time step conditioned on a visual vector, a previous hidden state, and a previously generated word.

Loss. During training, we adopt multiple losses to optimize the whole framework and effectively highlight the optimal graph nodes for manipulating attributes, relations, or objects in each modality with a global perspective. Overall, two commonly used loss functions in generation tasks, namely the GAN loss L_{GAN} and the Encoder-Decoder loss L_{ED}, as well as the designed cross-modal loss L_{CM} are adopted. Significantly, we introduce the cross-modal loss L_{CM} which further helps to shape the feature space:

$$L_H = \frac{1}{\left\|\left|H_{\mathbf{G}_{I(,S)}}\right| - \left|H_{I(,S)}\right|\right\|} + WS\left[\left(H_{\mathbf{G}_I}; H_I\right), \left(H_{\mathbf{G}_S}; H_S\right)\right],$$

$$L_M = \|M_I\|^2 + \|M_S\|^2 + \gamma \frac{1}{\||M_I| - |M_S|\|}.$$

We have $L_{CM} = L_H + L_M$, where $|\cdot|$ indicates the first norm of the matrix, and $WS[a, b]$ computes the Wasserstein distance between a and b. For L_H, the first term encourages the bridge nodes to sufficiently learn the representation from other modalities, while the second term constrains the joint distribution of the representation from both modalities to stay

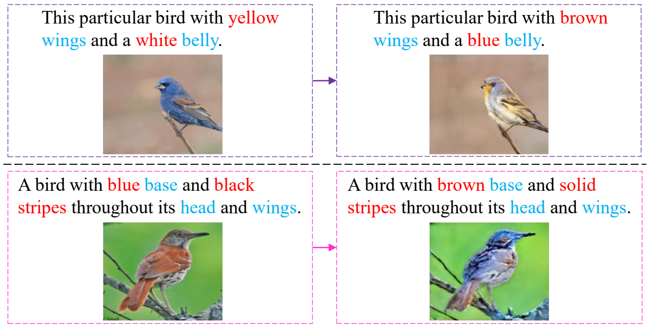

This particular bird with yellow wings and a white belly.
This particular bird with brown wings and a blue belly.
A bird with blue base and black stripes throughout its head and wings.
A bird with brown base and solid stripes throughout its head and wings.

Figure 4: Qualitative results on the CUB dataset. We visualize the color manipulation in the first row, and the texture manipulation is shown in the second row.

Method	BLEU	TR_s	LD_r	Top-1	AP@50
AttnGAN	-	-	-	0.55	0.51
TaGAN	-	-	-	0.61	0.63
ManiGAN	-	-	-	0.72	0.70
Ours	**36.1**	**0.74**	**0.83**	**0.79**	**0.75**

Table 3: Quantitative results on the CUB dataset, where all indicators return the average value. We do not report the image-text similarity scores for other methods which can not generate sentences.

close, maintaining the stability of the learning process. For L_M, The first two items are adopted to force M_I and M_S to be sparse so that only a portion of cross-modal nodes are highlighted for the graph-based aggregation. The third term with the weighting factor γ makes the highlighted nodes by M_I, and M_S tend to be distinctive after the relation reasoning process, thus facilitating the exchange of cross-modal nodes and avoiding the overfitting during training. The final loss becomes: $L = L_{GAN} + L_{ED} + L_{CM}$.

4 Experiments

Experimental Setup. We verified our baseline in two aspects, including the manipulation of both attributes and relations of objects. For cross-modal attribute manipulation, we evaluated our method on the CUB dataset [Wah *et al.*, 2011]. For cross-modal relation manipulation, we validated on the Visual Genome (VG) [Krishna *et al.*, 2017].

Quantitative Results. The quantitative evaluations are conducted on both synthesized images and generated sentences. For synthesized images, we conducted a human evaluation and compared to relevant generative models including SISGAN [Dong *et al.*, 2017], AttnGAN [Xu *et al.*, 2018], TaGAN [Nam *et al.*, 2018], ManiGAN [Li *et al.*, 2020a], and Lightweight GAN [Li *et al.*, 2020b].

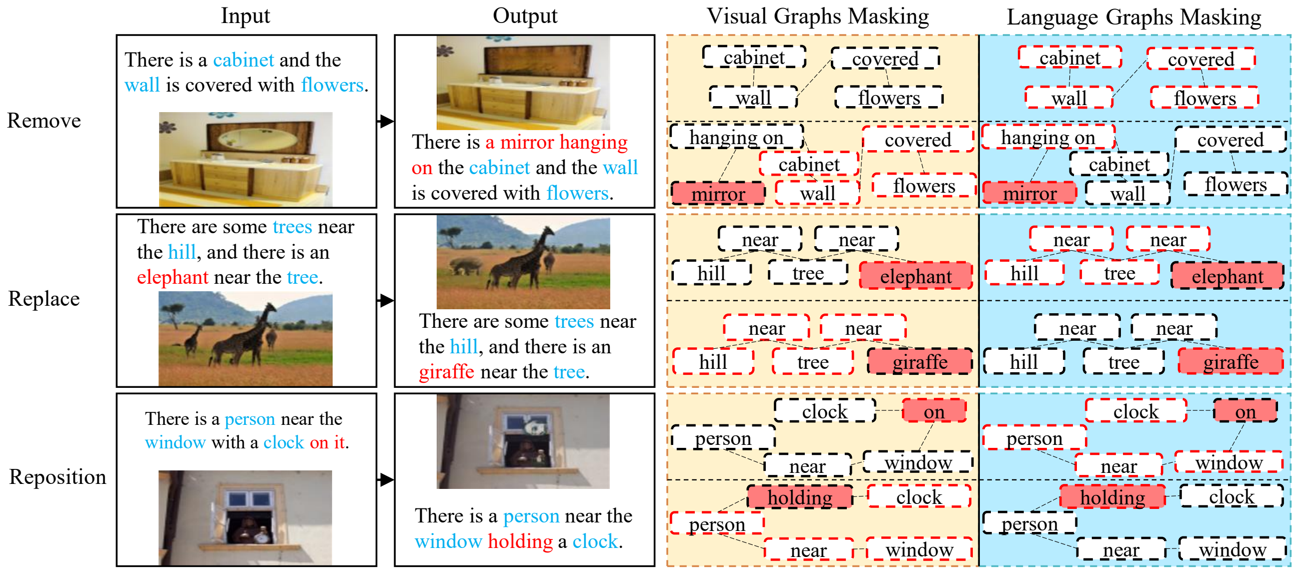
We randomly selected 20 images and 10 texts from the test set of the CUB dataset and produced 200 image-sentence pairs for each method. We invited 80 workers to compare the results after looking at the input image, the input sentence, and both outputs based on two criteria: (i) Accuracy: whether the visual attributes (colors, textures) of the manipulated image match the text, and the background-independent to the text is preserved, and (ii) Naturalness: whether the manipulated image looks natural and visually pleasing. Both criteria are categorized into three ranks (1, 2, and 3), and the lower, the better. We collected a total of 16000 results. Besides, we computed $L2$ reconstruction error by forwarding images with the ground-truth text descriptions, while Inception Score (IS) and Fréchet Inception Distance (FID) are also evaluated on a large number of modified samples produced from mismatched pairs, i.e., randomly chosen input images edited by randomly selected text descriptions.

As shown in Table 1, our method consistently achieves the highest average ranking on both Accuracy and Naturalness, the lowest reconstruction error, and better IS and FID values. It indicates that our method generates more realistic images, where the visual attributes are manipulated accurately with the given descriptions and effectively preserve the text-independent contents. We computed the parameters and recorded the runtime for training a single optimization epoch (RPE) and the inference time (IT) for generating 100 new modified images to evaluate the efficiency. The results indicate that our method is more friendly to memory-limited devices.

In addition, we evaluated our method on the Visual Genome to verify its effectiveness in generating natural images. Since there is no ground truth for manipulations, we formulated the quantitative evaluation as image reconstruction. In this case, we manipulated the relations between objects from the language descriptions and measured the reconstruction quality given an input image. Table 2 shows the reconstruction errors of comparative methods, showing that our method significantly outperforms Cond-sg2im [Johnson *et al.*, 2018] and SIMSG [Dhamo *et al.*, 2020]. Notably, our method achieves better results on all frequently-used reconstruction metrics (MAE, SSIM, LPIPS) [Dhamo *et al.*, 2020] and dominates for both inception score and FID, indicating higher visual quality. Compared to SIMSG, which changes the relations among objects by manipulating the generated scene graphs, our method provides a way to directly manipulate the image by inputting simple language descriptions that contain the desired modifications, improving efficiency with less manual effort from the user.

For generated sentences, BLEU [Papineni *et al.*, 2002] score, the TextRank criteria (TR_s) [Mihalcea and Tarau, 2004], and the Levenstein distance ratio (LD_r) [Yujian and Bo, 2007] are used for measuring the image-text similarity scores of our method. Furthermore, we compared the top-1 image-to-text retrieval accuracy (Top-1), and the percentage of the matching images in the top-50 text-to-image retrieval results (AP@50) with AttnGAN, TaGAN, and ManiGAN on the CUB dataset. As shown in Table 3, our proposal outperforms other methods on the image-text matching task, fully maintaining the image-independent contents from the input sentences to guide the sentence generation.

Qualitative Results. Figure 4 shows the qualitative results for cross-modal attribute manipulations on the CUB dataset, verifying that the attributes of visual modality are accurately exchanged with the corresponding attributes in language modality. This indicates our method: (1) effectively



There is a cabinet and the wall is covered with flowers.

There is a mirror hanging on the cabinet and the wall is covered with flowers.

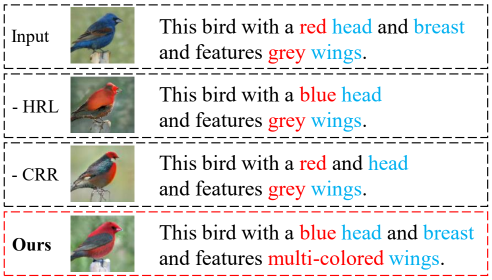
There are some trees near the hill, and there is an elephant near the tree.

There are some trees near the hill, and there is an giraffe near the tree.

There is a person near the window with a clock on it.

There is a person near the window holding a clock.

Figure 5: Qualitative results on the Visual Genome. For cross-modal relation reasoning, nodes with black borders are masked, while nodes with the red borders will be used as active nodes for subsequent processing. Nodes with the red fill indicate the focused objects or relations.



This bird with a red head and breast and features grey wings.

This bird with a blue head and features grey wings.

This bird with a red and head and features grey wings.

This bird with a blue head and breast and features multi-colored wings.

Figure 6: Qualitative ablation on the CUB dataset. (- HRL): removes the Heterogeneous Representation Learning module, and (- CRR): removes the Cross-modal Relation Reasoning module.

disentangles visual attributes invariant to pose, shape, and background in the visual modality; explicitly locating and exchanging the focused attributes with the language modality to reconstruct for obtaining images with detailed textures; (2) efficiently captures linguistic attributes by sufficiently analyzing the syntactic structure and content of the text in the language modality; accurately locating and exchanging the focused attributes with the visual modality to produce high-quality sentences.

To verify the effectiveness of our method for cross-modal relation manipulations, we configured three different settings, including object *removal*, *replacement*, and *reposition*, and evaluated on the Visual Genome. All the image manipulations are controlled by changing the corresponding text descriptions. As shown in Figure 5, Our method successfully removes the focused target (*mirror*) in the synthesized image based on the text description, while text-independent visual features from the input image can be successfully transferred to the output. In object *replacement*, we provided diverse replacements from foreground objects to background components. In Figure 5, our method removes the old object (‘*giraffe*’) very naturally from the input image while the substituted object (‘*elephant*’) adapts well to the synthesized image. A more challenging scenario is to change the relations among objects, which typically involves object *reposition*. The result in Figure 5 verifies that our method can differentiate between semantic concepts such as ‘*on*’ and ‘*holding*’, and the focused object (‘*clock*’) are re-arranged meaningfully according to the relation change indicated in the input sentence. In addition to the relation manipulations of focused objects, our method performs well for the background with uniform texture and complex structures. Moreover, the generated sentences are high-quality and meaningful, and the exchanges of target relations with the input images are effectively accomplished.

Components Ablation. We ablated the major components of our method qualitatively in Figure 6. First, removing the HRL module is not conducive to cross-modal representation learning for semantic alignment, leading to difficulties in implementing transformations of both visual and linguistic information from the inputs to the outputs; second, removing the CRR module damages the bidirectional information control in cross-modal relation reasoning, generating uncontrolled and unsatisfactory results on both modalities.

5 Conclusion

We have presented a strong baseline for cross-modal graph representation learning and relation reasoning and applied our proposal to the novel cross-modal adaptive manipulation task. Experimental results verify the superior performance of our method in cross-modal applications. We leave it for future work to extend our research to more modalities.

Acknowledgements

This work was supported by NSFC project Grant No. U1833101, SZSTC Grant No. JCYJ20190809172201639, ZDSYS20210623092001004 and WDZC20200820200655001.

References

[Cornia *et al.*, 2019] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8307–8316, 2019.

[Dhamo *et al.*, 2020] Helisa Dhamo, Azade Farshad, Iro Laina, Nassir Navab, Gregory D Hager, Federico Tombari, and Christian Rupprecht. Semantic image manipulation using scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5213–5222, 2020.

[Dong *et al.*, 2017] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5706–5714, 2017.

[Goodfellow *et al.*, 2020] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[Johnson *et al.*, 2018] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018.

[Kampffmeyer *et al.*, 2019] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P Xing. Rethinking knowledge graph propagation for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11487–11496, 2019.

[Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[Krishna *et al.*, 2017] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[Li *et al.*, 2019] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Controllable text-to-image generation. *arXiv preprint arXiv:1909.07083*, 2019.

[Li *et al.*, 2020a] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020.

[Li *et al.*, 2020b] Bowen Li, Xiaojuan Qi, Philip HS Torr, and Thomas Lukasiewicz. Lightweight generative adversarial networks for text-guided image manipulation. *arXiv preprint arXiv:2010.12136*, 2020.

[Liu *et al.*, 2019] Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. Fine-grained fact verification with kernel graph attention network. *arXiv preprint arXiv:1910.09796*, 2019.

[Mihalcea and Tarau, 2004] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[Nam *et al.*, 2018] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: Manipulating images with natural language. *arXiv preprint arXiv:1810.11919*, 2018.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*, 2011.

[Wang *et al.*, 2018] Zhouxia Wang, Tianshui Chen, Jimmy Ren, Weihao Yu, Hui Cheng, and Liang Lin. Deep reasoning with knowledge graph for social relationship understanding. *arXiv preprint arXiv:1807.00504*, 2018.

[Xu *et al.*, 2018] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[Yao *et al.*, 2019] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377, 2019.

[Yujian and Bo, 2007] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095, 2007.