CVPR JUNE 3-7, 2026 DENVER COLORADO

# STiTch : Semantic Transition and Transportation in Collaboration for Training-Free Zero-Shot Composed Image Retrieval

Miaoge Li, Dongsheng Wang, Zening Sun, Jinsen Zhang, Wenhan Luo, Jingcai Guo
The Hong Kong Polytechnic University, Hong Kong SAR & Shenzhen University, Shenzhen, China

# Problem Definition

<Reference Image>

*Add mountains in the background and change the number of dogs to nine.*

<Text Modification>

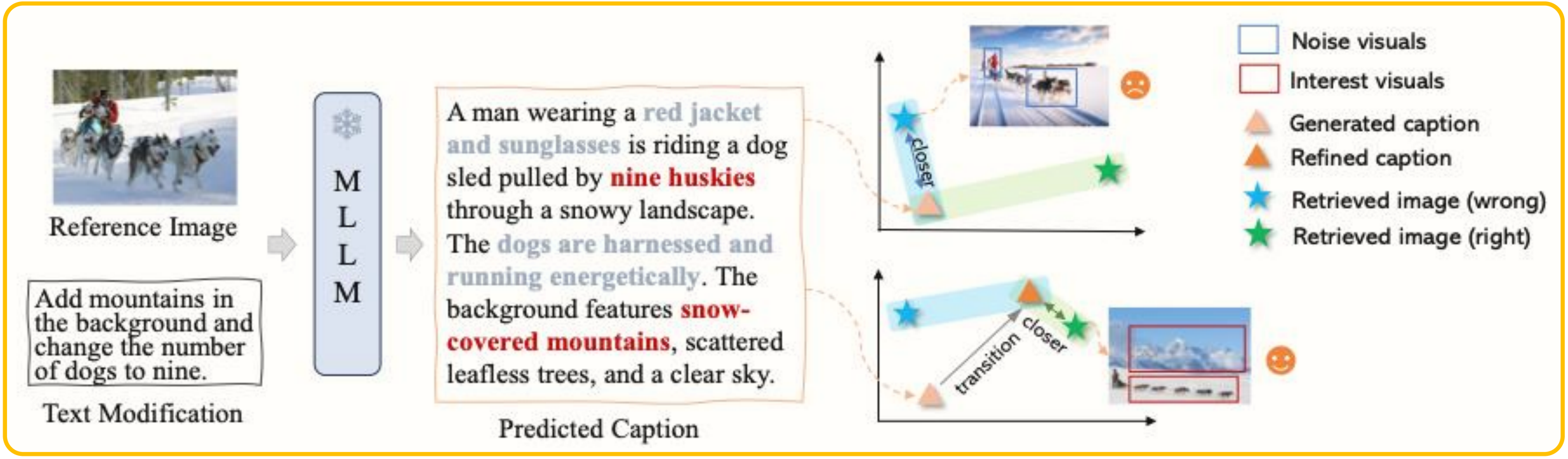
<Reference Image>

...

Retrieval Image Database

Composed Image Retrieval (CIR) :
Search a target image from an image database that **satisfies the semantic consistency with both reference image and text modification**.

# Motivation




***Extraneous Cognitive Load***: The reference image may trigger information leakage, which in turn leads to overemphasis on irrelevant details, affecting the retrieval performance.

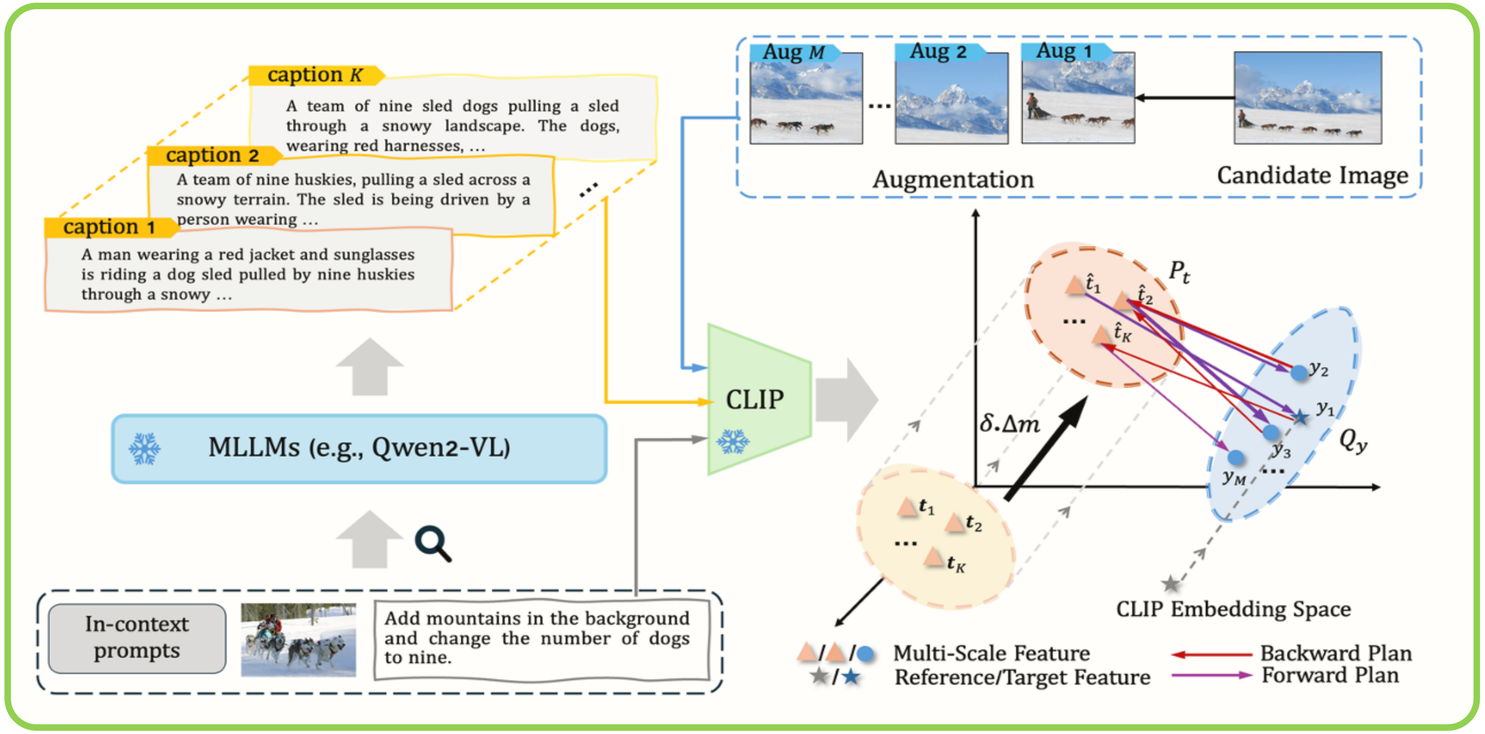
***Weak Compositional Modeling***: Point-to-point alignment during the retrieval stage fails to capture diverse compositions, leading to semantic imbalance across modalities.

# Our Proposed STiTch

## ① *Query*

In-context learning of MLLM

Provide a comprehensive understanding of the given query input.




## ② *Semantic Transition*

$$\Delta \boldsymbol{m} = \boldsymbol{y} - \boldsymbol{x}, \Delta \hat{\boldsymbol{m}} = f(m)$$
$$\hat{\boldsymbol{t}}_k = (1-\alpha)\boldsymbol{t}_k + \alpha \Delta \hat{\boldsymbol{m}}$$

Highlight semantic diversity while avoiding unnecessary information from the reference image.

## ③ *Transportation*

$$P_t = \frac{1}{K}\sum_{k=1}^{K}\delta_{\boldsymbol{t}_k} \quad Q_y = \frac{1}{M}\sum_{m=1}^{M}\delta_{\boldsymbol{y}_m}$$

$$\begin{aligned}\mathcal{L}_{bi}(P_t, Q_y) &= \mathcal{L}_{P_t \to Q_y}(P_t, Q_y) + \mathcal{L}_{Q_y \to P_t}(P_t, Q_y) \\ &= \sum_{m,k} \pi(\boldsymbol{y}_m|\hat{\boldsymbol{t}}_k)c(\hat{\boldsymbol{t}}_k, \boldsymbol{y}_m) + \pi(\hat{\boldsymbol{t}}_k|\boldsymbol{y}_m)c(\boldsymbol{y}_m, \hat{\boldsymbol{t}}_k)\end{aligned}$$

Develop a bidirectional transportation distance to achieve fine-grained set-to-set alignment between modalities.

# Experimental Results

| CIRCO + CIRR → | | | CIRCO | | | | CIRR | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Metric | | mAP@k | | | | Recall@k | | | $\text{Recall}_{\text{Subset}}$@k | | |
| Arch | Method | Train | k=5 | k=10 | k=25 | k=50 | k=1 | k=5 | k=10 | k=1 | k=2 | k=3 |
| ViT-L/14 | Pic2Word | ✓ | 8.72 | 9.51 | 10.64 | 11.29 | 23.90 | 51.70 | 65.30 | 53.76 | 74.46 | 87.08 |
| | SEARLE | ✓ | 11.68 | 12.73 | 14.33 | 15.12 | 24.24 | 52.48 | 66.29 | 53.76 | 75.01 | 88.19 |
| | LinCIR | ✓ | 12.59 | 13.58 | 15.00 | 15.85 | 25.04 | 53.25 | 66.68 | 57.11 | 77.37 | 88.89 |
| | Context-I2W | ✓ | 13.04 | 14.62 | 16.14 | 17.16 | 25.60 | 55.10 | 68.50 | - | - | - |
| | CIReVL | ✗ | 18.57 | 19.01 | 20.89 | 21.80 | 24.55 | 52.31 | 64.92 | 59.54 | 79.88 | 89.69 |
| | LDRE | ✗ | 23.35 | 24.03 | 26.44 | 27.50 | 26.53 | 55.57 | 67.54 | 60.43 | 80.31 | 89.90 |
| | OSrCIR | ✗ | 23.87 | 25.33 | 27.84 | 28.97 | **29.45** | 57.68 | 69.86 | 62.12 | 81.92 | 91.10 |
| | SEIZE | ✗ | 24.98 | 25.82 | 28.24 | 28.35 | 28.65 | 57.16 | 69.23 | 62.22 | 84.05 | 92.34 |
| | **STiTch(Ours)** | ✗ | **25.55** | **26.27** | **28.81** | **29.99** | 28.87 | **57.97** | **69.90** | **65.22** | **84.10** | **92.37** |
| ViT-G/14 | CIReVL | ✗ | 26.77 | 27.59 | 29.96 | 31.03 | 34.65 | 64.29 | 75.06 | 67.95 | 84.87 | 93.21 |
| | LDRE | ✗ | 31.12 | 32.24 | 34.95 | 36.03 | 36.15 | 66.39 | 77.25 | 68.82 | 85.66 | 93.76 |
| | OSrCIR | ✗ | 30.47 | 31.14 | 35.03 | 36.59 | 37.26 | 67.25 | 77.33 | 69.22 | 85.28 | 93.55 |
| | SEIZE | ✗ | 32.46 | 33.77 | 36.46 | 37.55 | 38.87 | 69.42 | 79.42 | **74.15** | 89.23 | 95.71 |
| | **STiTch (Ours)** | ✗ | **34.40** | **35.56** | **38.07** | **40.02** | **39.23** | **69.95** | **79.56** | 73.56 | **89.50** | **95.86** |

| GeneCIS → | | | Focus Attribute | | | Change Attribute | | | Focus Object | | | Change Object | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Arch | Method | Train | R@1 | R@2 | R@3 | R@1 | R@2 | R@3 | R@1 | R@2 | R@3 | R@1 | R@2 | R@3 | R@1 |
| ViT-L/14 | SEARLE | ✓ | 17.1 | 29.6 | 40.7 | 16.3 | 25.2 | 34.2 | 12.0 | 22.2 | 30.9 | 12.0 | 24.1 | 33.9 | 14.4 |
| | LinCIR | ✓ | 16.9 | 30.0 | 41.5 | 16.2 | 28.0 | 36.8 | 8.3 | 17.4 | 26.2 | 7.4 | 15.7 | 25.0 | 12.2 |
| | Context-I2W | ✓ | 17.2 | 30.5 | 41.7 | 16.4 | 28.3 | 37.1 | 8.7 | 17.9 | 26.9 | 7.7 | 16.0 | 25.4 | 12.7 |
| | CIReV | ✗ | 19.5 | 31.8 | 42.0 | 14.4 | 26.0 | 35.2 | 12.3 | 21.8 | 30.5 | 17.2 | 28.9 | 37.6 | 15.9 |
| | OSrCIR | ✗ | **20.9** | 33.1 | 44.5 | 17.2 | 28.5 | 37.9 | 15.0 | 23.6 | 34.2 | 18.4 | 30.6 | 38.3 | 17.9 |
| | SEIZE | ✗ | 20.5 | 33.4 | 45.0 | 17.6 | 28.9 | 38.5 | 15.4 | 25.6 | 36.2 | 18.7 | 30.9 | 39.8 | 18.1 |
| | **STiTch(Ours)** | ✗ | 20.3 | **34.6** | **46.4** | **18.3** | **29.8** | **41.6** | **16.8** | **28.5** | **38.4** | **18.8** | **31.0** | **40.3** | **18.6** |
| ViT-G/14 | CIReVL | ✗ | 20.9 | 34.4 | 44.9 | 16.5 | 29.0 | 39.8 | 15.1 | 25.6 | 33.4 | 18.5 | 31.6 | 41.4 | 17.8 |
| | OSrCIR | ✗ | 22.7 | 36.4 | 47.0 | 17.9 | 30.8 | 42.0 | 16.9 | 28.4 | 36.7 | **21.0** | **33.4** | **44.2** | 19.6 |
| | SEIZE | ✗ | **22.9** | 36.2 | 47.3 | 18.6 | 31.4 | 42.7 | 18.2 | 28.8 | 37.6 | 19.6 | 33.0 | 43.5 | 19.8 |
| | **STiTch (Ours)** | ✗ | 21.9 | **36.4** | **47.9** | **19.6** | **31.9** | **42.8** | **20.2** | **30.3** | **39.6** | 19.7 | 33.2 | 43.4 | **20.4** |

# Ablation Study

## *Main component analysis*

| CIRCO + CIRR → | | CIRCO | | | | CIRR | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Strategy | | mAP@k | | | | Recall@k | | | $\text{Recall}_{\text{Subset}}$@k | | |
| *Transition* | *Transportation* | k=5 | k=10 | k=25 | k=50 | k=1 | k=5 | k=10 | k=1 | k=2 | k=3 |
| ✗ | ✗ | 31.23 | 32.87 | 36.32 | 38.04 | 37.22 | 67.36 | 77.84 | 69.93 | 86.48 | 94.05 |
| ✓ | ✗ | 31.89 | 34.46 | 37.94 | 39.67 | 38.33 | 68.45 | 78.03 | 72.81 | 88.13 | 94.51 |
| ✗ | ✓ | 32.14 | 34.78 | 37.87 | 39.48 | 38.48 | 68.38 | 78.34 | 72.15 | 88.04 | 94.48 |
| ✓ | ✓ | **34.40** | **35.56** | **38.07** | **40.02** | **39.23** | **69.95** | **79.56** | **73.56** | **89.50** | **95.86** |

## *Impacts of caption number and augmentation views*

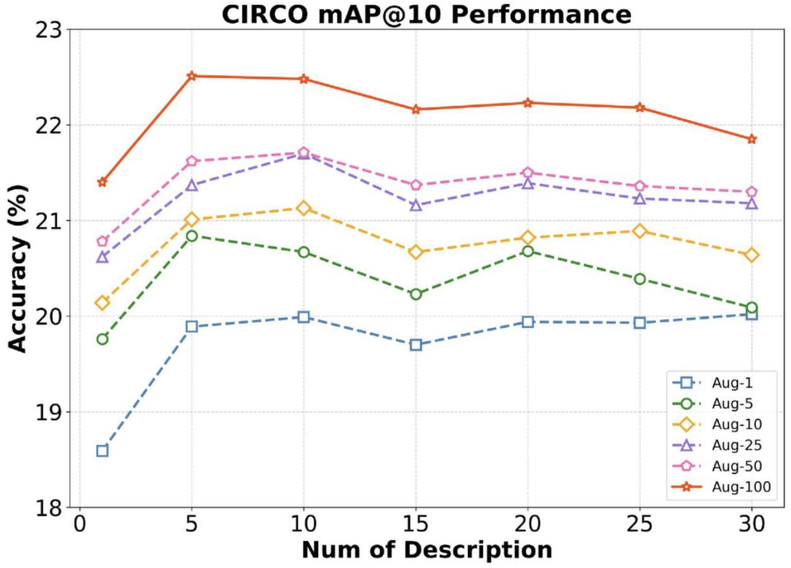



Figure 1. Ablation on the number of descriptions and image augmentations (Aug-10: 10 augmentations per image) on the CIRCO dataset with CLIP-B/32.

## *Impacts of the bidirectional distance*

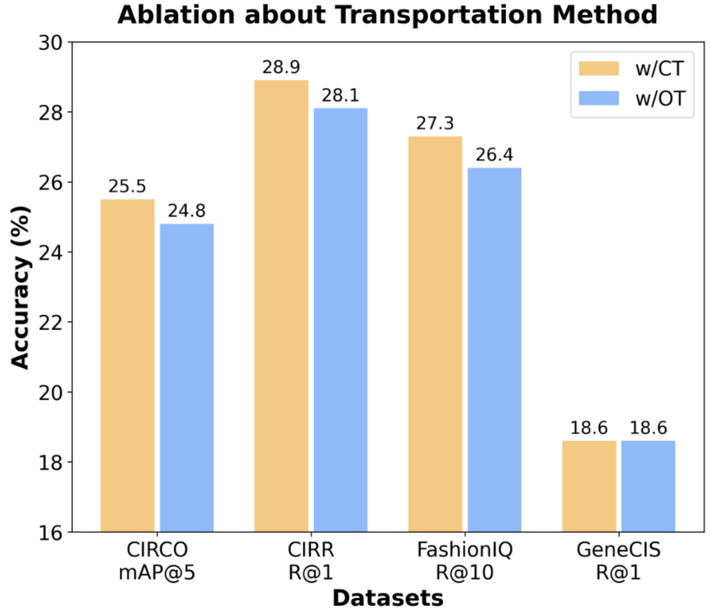



Figure 2. Ablation results of different alignment strategies across four datasets with CLIP-L/14.

# Visualization Analysis

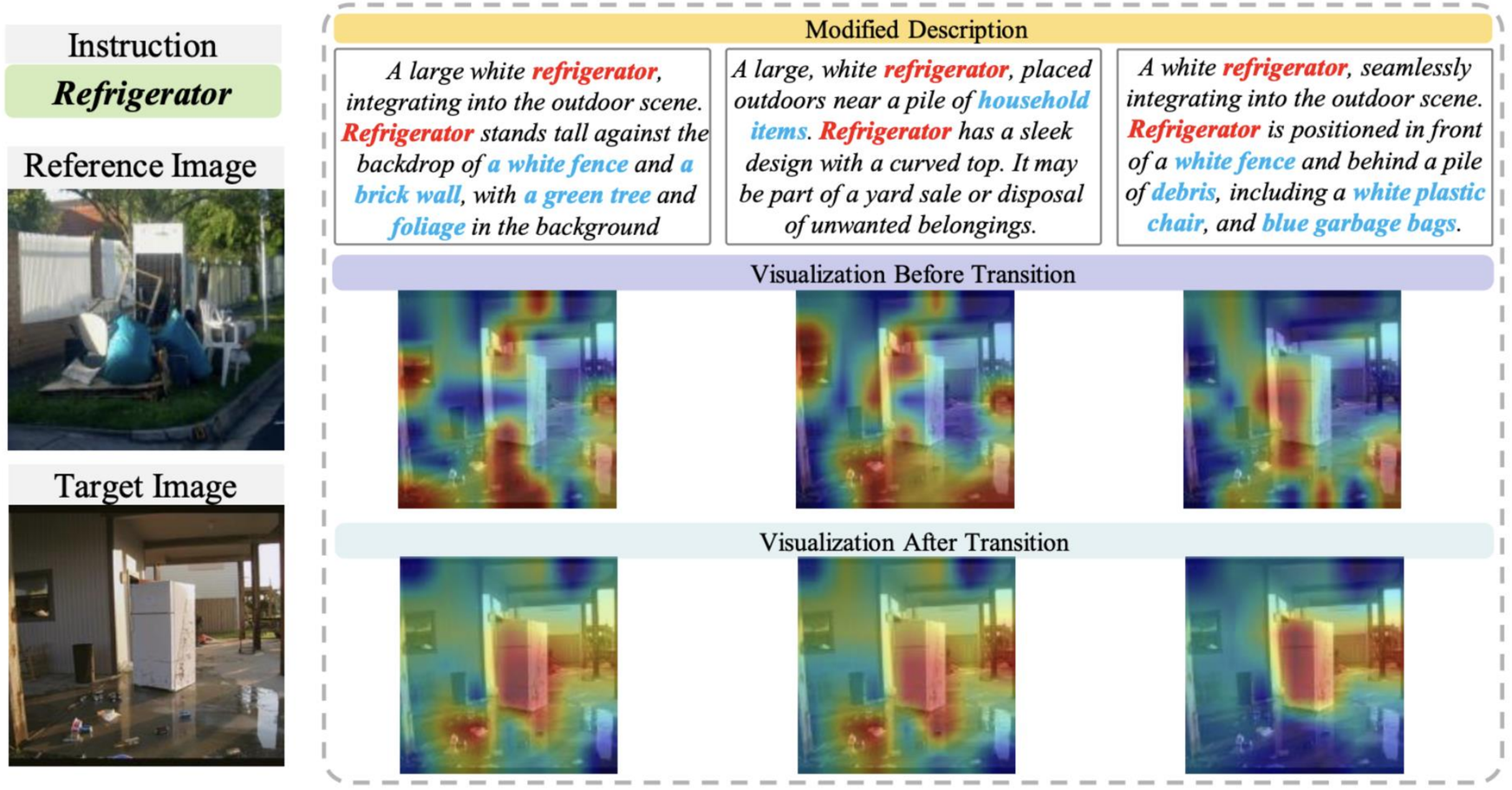



Figure 3. Visualization of the GeneCIS dataset on the 'Focus Object' task. Heatmaps before and after the transition on target image are shown.

Captions generated by MLLMs often contain irrelevant visual noise (***blue*** text), while the STiTch model effectively suppresses such noise and highlights the correct focus object (***red*** text).