LQMFormer: Language-aware Query Mask Transformer for Referring Image Segmentation

Supplementary Material

1. Gaussian Enhanced Multi-Modal Fusion Module (GMMF)

The application of the discrete Fast Fourier Transform (FFT) within the GMMF module is central to transforming the vision-language features $F_{vl}^{'}$ into the frequency domain, for efficient global modelling. The FFT is defined as:

$$F_{\text{FFT}}(F_{vl}^{'}) = \mathcal{F}(F_{vl}^{'}) \quad (1)$$

where $\mathcal{F}$ denotes the Fourier transform operation, and $F_{vl}^{'}$ are the vision-language features that are obtained from vision backbone model.

Amplitude and Phase Components in Fourier Transform

The complex exponential in $\mathcal{F}(F_{vl}^{'})$ contains the phase information. The amplitude $\mathcal{A}(F_{vl}^{'})$ and phase $\mathcal{P}(F_{vl}^{'})$ components, critical for reconstructing and manipulating the vision-language features in the frequency domain, are derived from $\mathcal{F}(F_{vl}^{'})$ and defined as follows:

$$\mathcal{A}(F_{vl}^{'}) = \sqrt{\mathcal{R}(F_{vl}^{'})^2 + \mathcal{I}(F_{vl}^{'})^2}, \quad (2)$$

$$\mathcal{P}(F_{vl}^{'}) = \arctan\left[\frac{\mathcal{I}(F_{vl}^{'})}{\mathcal{R}(F_{vl}^{'})}\right], \quad (3)$$

where $\mathcal{R}(F_{vl}^{'})$ and $\mathcal{I}(F_{vl}^{'})$ represent the real and imaginary components of $\mathcal{F}(F_{vl}^{'})$, respectively.

Amplitude Modulation

In the GMMF module, the amplitude component $\mathcal{A}(F_{vl}^{'})$ is modulated to enhance global visual-language features, specifically using Gaussian smoothed filters for low-pass filtering [1, 3]:

$$\mathcal{A}'(F_{vl}^{'}) = \mathcal{A}(F_{vl}^{'}) * \phi(F_{vl}^{'}, \beta), \quad (4)$$

where $*$ represents the operation of low-pass filtering, and the result $\mathcal{A}'(F_{vl}^{'})$ is the low-pass filtered version of the amplitude component and β is learnable bandwidth parameter, calculated from F_{vl} followed by sequence of Linear and Pooling operation.

Now, for reconstructing the complex frequency representation from the amplitude and phase, the following equation is used:

$$\mathcal{F}'(F_{vl}^{'}) = \mathcal{A}'(F_{vl}^{'}) * e^{j\mathcal{P}(F_{vl}^{'})}, \quad (5)$$

where $e^{j\mathcal{P}(F_{vl}^{'})}$ represents the complex exponential with the phase component $\mathcal{P}(F_{vl}^{'})$.

To reconstruct the enhanced features for further processing, following [3], we apply the inverse discrete Fast Fourier transform (Inverse FFT):

$$F_{vl} = F_{\text{FFT}}^{-1}(\text{Conv}(\mathcal{F}'(F_{vl}^{'}))) + F_{vl}^{'}, \quad (6)$$

where F_{FFT}^{-1} denotes the inverse Fourier transform operation, and Conv indicates 1x1 convolution. The final step in the GMMF module involves the combination of the original features $F_{vl}^{'}$ and the Gaussian-enhanced features, here represented by F_{vl}, resulting from the convolution and inverse FFT of the modulated complex frequency representation $\mathcal{F}'(F_{vl}^{'})$.

This reconstruction is critical in the GMMF module, particularly for modulating the amplitude component while preserving the phase information, leading to enhanced visual-language feature representation. This detailed section supplements method section by offering a more detailed explanation of the operations within the GMMF module, particularly focusing on the Fourier transform applications and the rationale behind the use of Gaussian smoothing in the frequency domain for feature enhancement.

2. Qualitative Results

The qualitative analysis shows the LQMFormer capabilities in segmenting objects from complex visual scenes based on intricate language descriptions. Each row in Figure 1 reflects a different scenario where the model proficiency in vision grounding and language understanding is evaluated.

Contextual Differentiation: The first row illustrates the model's ability to distinguish between individuals based on their roles and attire – a critical skill in scenes with multiple similar subjects. Notably, it segments the batter in white and the referee in black, showing understanding of context based description.

Detailed Descriptive Segmentation: The second row presents a case where LQMFormer precisely segments a batter based on both a descriptive action ("with one knee on the ground") and a specific object relation ("a bat on the ground in front of him"). This shows the model's capacity to interpret complex activities and relative positioning.

Disambiguation of Multiple Entities: In the third row, LQMFormer adeptly segments both a man and a woman wearing purple, accurately understanding multiple entities

Figure 1. Qualitave Comparison of our model LQMFormer with ReLA [2] on GRES dataset.

based on their clothes color, which is a common challenge in crowded scenes.

Understanding of Complex Indirect Descriptions: The fourth row demonstrates LQMFormer 's ability to understand indirect descriptors like ”crouching dude” in the foreground, effectively linking the alias to the complex description in the scene.

Challenging Case of Recognition of Occluded Objects: Conversely, the fifth row presents a challenging scenario for LQMFormer . The model encounters difficulty accurately segmenting an occluded person characterized solely by a color descriptor (”pink sweater”). This case shows the model's current limitations in distinguishing occluded objects, especially when they blend into the background, indicating a potential direction for future enhancements.

These qualitative evaluations shows that LQMFormer is proficient in scenarios that require detailed language understanding and vision grounding. The model understands a range of expressions, from straightforward references to a single entity to complex descriptions involving multiple objects. Its performance passes compared to previous methods in RIS, highlighting its capability in this complex task.

References

[1] Chongyi Li, Chun-Le Guo, Man Zhou, Zhexin Liang, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Embedding fourier for ultra-high-definition low-light image enhancement. *arXiv preprint arXiv:2302.11831*, 2023. 1

[2] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23592–23601, 2023. 2

[3] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 920–930, 2023. 1