Article:
Gao, M., Yang, J., Han, J. et al. (3 more authors) (2023) Decoupling multimodal transformers for referring video object segmentation. IEEE Transactions on Circuits and Systems for Video Technology, 33 (9). pp. 4518-4528. ISSN 1051-8215

https://doi.org/10.1109/TCSVT.2023.3284979

Decoupling Multimodal Transformers for Referring Video Object Segmentation

Mingqi Gao, Jinyu Yang, *Student Member, IEEE*, Jungong Han, *Member, IEEE*, Ke Lu, *Senior Member, IEEE*, Feng Zheng*, *Member, IEEE*, Giovanni Montana

***Abstract*—Referring Video Object Segmentation (RVOS) aims to segment the text-depicted object from video sequences. With excellent capabilities in long-range modelling and information interaction, transformers have been increasingly applied in existing RVOS architectures. To better leverage multimodal data, most efforts focus on the interaction between visual and textual features. However, they ignore the syntactic structures of the text during the interaction, where all textual components are intertwined, resulting in ambiguous vision-language alignment. In this paper, we improve the multimodal interaction by DECOUPLING the interweave. Specifically, we train a lightweight subject perceptron, which extracts the subject part from the input text. Then, the subject and text features are fed into two parallel branches to interact with visual features. This enables us to perform subject-aware and context-aware interactions, respectively, thus encouraging more explicit and discriminative feature embedding and alignment. Moreover, we find the decoupled architecture also facilitates incorporating the vision-language pre-trained alignment into RVOS, further improving the segmentation performance. Experimental results on all RVOS benchmark datasets demonstrate the superiority of our proposed method over the state-of-the-arts. The code of our method is available at: https://github.com/gaomingqi/dmformer.**

***Index Terms*—Decoupled multimodal transformers, Referring video object segmentation, Vision-language pre-training.**

I. Introduction

Segmenting the object of interest in videos is a fundamental procedure in video analysis and editing [1], thus has spawned several relevant tasks and attracted much attention in the community. Based on the ways specifying the target objects, existing Video Object Segmentation (VOS) approaches can be mainly categorised into three folds: semi-supervised (also termed as one-shot) VOS, interactive VOS, and referring VOS (RVOS). The first two approaches rely on the target objects specified by human annotations [2]. In contrast, the object of interest in RVOS is described by language expressions. This renders RVOS more user-friendly and compatible with human-computer interaction applications.

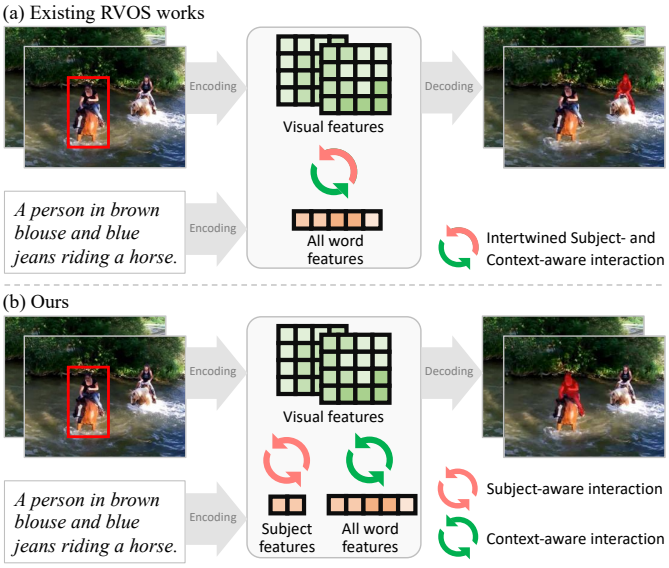

Fig. 1. Comparison of the RVOS works with different interaction strategies. Given an input video and a text referring to the target object, (a) intertwines all textual components to interact with visual features; (b) decouples the interaction into subject-aware and context-aware, based on subject perception from the text. Such decoupling encourages more discriminative and comprehensive vision-language interaction. Bounding boxes in the left column indicate the ground truth objects. Red masks in the right column are the predictions. Best viewed in colour.

Apparently, RVOS is essentially under the multimodal setting, in which the key challenge is how to effectively interact between vision and language features. Existing methods achieve this via dynamic convolution [3], [4], crossmodal attention [5]–[13], or transformers [14]–[19], making remarkable progress. Theoretically, these methods are designed to attend to all multimodal elements and perform comprehensive interactions, which, however, is hard to achieve in practice. This is mainly because they ignore the syntactic structures of the text and intertwine all textual components during the interaction. In most training samples, the text-referred object is the only object in its category such that RVOS methods tend to focus more on the subject rather than other valuable descriptions in the text. Therefore, explicit vision-language

*Corresponding author.

This work was supported by the National Key R&D Program of China (Grant NO. 2022YFF1202903) and the National Natural Science Foundation of China (Grant NO. 62122035 and 61972188).

Mingqi Gao is with Southern University of Science and Technology, Shenzhen 518055, China, and also with University of Warwick, Coventry CV1 7AL, U.K. (e-mail: mingqi.gao@outlook.com).

Jinyu Yang is with Southern University of Science and Technology, Shenzhen 518055, China, and also with University of Birmingham, Birmingham B15 2TT, U.K. (e-mail: jinyu.yang96@outlook.com).

Jungong Han is with University of Sheffield, Sheffield S10 2TN, U.K., and also with University of Warwick, Coventry CV1 7AL, U.K. (e-mail: jungonghan77@gmail.com).

Ke Lu is with University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: luk@ucas.ac.cn).

Feng Zheng is with Southern University of Science and Technology, Shenzhen 518055, China (e-mail: f.zheng@ieee.org).

Giovanni Montana is with University of Warwick, Coventry CV4 7AL, U.K. (e-mail: g.montana@warwick.ac.uk).

Manuscript received April 19, 2021; revised August 16, 2021.

interaction remains unexplored to current RVOS works, thus performing unsatisfactorily in practical scenarios where multiple distractors and complex scenes coexist.

To tackle the challenge, our idea is to DECOUPLE multi-modal interactions. As shown in Fig. 1, given a video and a text referring to the target, we first locate the subject part from the text, via a lightweight subject perceptron. Then, with the separation, two parallel multimodal interactions are performed: Subject-aware interaction and Context-aware interaction. The former focuses on subject objects, and the latter is encouraged to mine discriminative features from the description part of the text. Unlike existing RVOS methods, which interact visual features with all textual elements, our decoupled strategy assigns explicit tasks to different interaction branches. Therefore, the excessive focus can be effectively mitigated to achieve discriminative feature embedding and interaction.

As the foundation for multimodal interaction, feature alignment between different modalities is also critical to RVOS. Most existing works achieve this implicitly by minimising the segmentation loss on RVOS datasets [3], [20], [21]. However, limited by annotation efforts, these datasets cannot provide sufficient video-text pairs for alignment learning. In addition, such insufficiency significantly restricts the RVOS methods' generalisability to the visual or textual concepts that are rare/unseen in the training set.

Our method improves the alignment by incorporating the vision-language pre-trained knowledge. Recently, Vision-Language Pre-training (VLP) has boosted many vision and multi-modal applications due to its rich semantic correspondence and impressive transferability [22]. However, such advances have not yet been explored in RVOS since there are differences in targets between current VLP works (image-level [23] or object-level [24]) and RVOS (pixel-level), which hinders effective knowledge transfer. This paper makes the first attempt to address this challenge. Specifically, we incorporate the knowledge from an object-level VLP model [24] to boost our multi-modal encoder. Despite our ultimate goal being pixel-level prediction, we found the proposed decoupling mechanism coincidentally enables part of our architecture (subject-aware interaction) to have a similar purpose as the VLP model [24]. This way, our architecture can better benefit from VLP knowledge to align vision and language elements.

Our contributions can be summarised as follows:

- We propose a novel transformer-based RVOS architecture, termed Decoupled Multimodal Transformer (DMFormer), which explicitly interacts visual features with different syntactic components from the text. This would ultimately encourage more explicit and comprehensive feature interactions for RVOS.
- With the decoupled architecture, we explore an effective strategy to transfer knowledge from large-scale pre-trained vision-language alignment to RVOS.
- Our proposed DMFormer consistently outperforms the state-of-the-arts on all existing RVOS benchmarks, including Ref-YouTube-VOS [21], Ref-DAVIS [20], A2D-Sentence [3], and J-HMDB-Sentence [3].

II. Related Works

A. Referring Video Object Segmentation

Given the input video and text, the goal of RVOS is to segment and associate the text-referred object on all video frames. The task is firstly proposed by Gavrilyuk et al. [3] and further extended by Khoreva et al. [20] and Seo et al. [21] via broading the intput text from action-orient to unconstrained descriptions. To well link textual clues with visual objects, current RVOS methods mainly use three techniques to interact between multimodal elements: dynamic convolution, cross-modal attention, and transformers.

Dynamic convolution is firstly applied to RVOS by Gavrilyuk et al. [3], where text features are encoded as kernels to convolve visual features. Wang et al. [4] improve this idea by modulating textual kernels based on the visual context to be convolved, bringing more robustness against visually similar distractors. Despite being effective, the information interaction via convolution is not sufficient to handle complex scenes and language expressions.

Crossmodal attention is a widely used technique in RVOS as it can build the fine-grained and semantic correspondence between vision and language elements. Earlier work [5] leverages such properties to refine visual contexts and reduce language variations. To utilise the text more sufficiently, several works incorporate specific language components into attention-based interaction. These components could be semantic words (e.g., entities, attributes, or relations) [6], [7] or the adaptively extracted text elements [8], [10]. More recently, attention-based RVOS works tend to take temporal information into account. For example, CSTM [9] and LBDT [12] interact language features with visual features from both intra- and inter-frames. MMVT [11] measures optical flow from adjacent frames and fuses flow maps with visual and language features. Besides temporal features, object-level features are also considered in a recent work [13], achieving multi-granularity and multi-modal attention for RVOS.

Transformers's success in both Natural Language Processing (NLP) [25] and Computer Vision (CV) [26] encourages relevant applications for multimodal analysis. Unlike the above works, transformer-based RVOS achieves vision-language interaction entirely based on the attention mechanism. Earlier methods [14]–[16] segment each video frame individually and only leverage transformers to fuse multimodal features. Inspired by the application of DETR [27], [28] in video instance segmentation [29], the recently proposed works resort to DETR-like architectures for RVOS [17]–[19]. Specifically, transformers are leveraged for multimodal feature fusion and object localisation across all video frames. Therefore, temporally consistent interaction and segmentation can be achieved in an end-to-end manner, which enables such design to outperform others in both accuracy and efficiency and forms a new baseline for future RVOS works.

Albeit achieving good performance, it is still challenging for existing RVOS works to perform comprehensive multimodal interaction. This is mainly because these works consider all textual components during the interaction. Since most training samples have no distractors with the same category as the

target object, the interaction is implicitly driven to focus more on the subject part of the text. As a result, current RVOS works are vulnerable to same-category backgrounds, as shown in Fig. 1 (a). In contrast, our method decouples multimodal interaction into two processes: subject-aware interaction and context-aware interaction, which can effectively raise more attention on the description part of the text, thereby enhancing the robustness of RVOS methods.

B. Vision-Language Pre-training

More recently, Vision-Language Pre-training (VLP) [22] has achieved significant advances and attracted much attention in the community. The basic idea of VLP is to learn vision-language feature alignment from large-scale image-text pairs. One of the most representative VLP methods is CLIP (Contrastive Language-Image Pre-training) [23], which collects 400M image-text pairs and formulates a contrastive objective to learn the alignment. The following-up experiments validate that the learned knowledge by CLIP has a strong transferability and can boost several downstream tasks such as visual captioning [30] and visual question answering [31].

Besides CLIP, there are several VLP works leveraging large-scale image-text pairs for vision-language alignment. The main difference between them lies in the data granularity. For example, CLIP encodes image-level and sentence-level representations. X-VLM [32] considers multi-grained visual concepts (ranging from object-level to image-level) to align with sentences. FILIP [33] and GLIP [24], [34] focus on patch-word and region-phrase alignment, respectively. All these works have shown their impressive capabilities to transfer to downstream tasks.

Despite VLP being popular in multimodal applications, it remains blank in RVOS, as there is a task gap between current VLP works and RVOS. Since the key to a successful transfer is the consistency between source and target tasks. Such a gap significantly blocks the knowledge flowing from VLP to RVOS. In this paper, we found the decoupled multimodal interaction implicitly reformulates RVOS closer to one of VLP works (GLIP [24]), whose knowledge could be incorporated into our proposed architecture seamlessly, enabling better vision-language alignment.

III. Method

We first present the overview of our Decoupled Multimodal Transformers (DMFormer) for RVOS in section III-A. In section III-B, we introduce how to perceive the subject from the input text. Then we illustrate the decoupled multimodal interaction and the incorporation of VLP knowledge in sections III-C and III-D, respectively. Finally, the optimisation and implementation details are provided in section III-E.

A. Overview

Fig. 2 illustrates the pipeline of our proposed Decoupled Multimodal Transformers (DMFormer) for RVOS. Analogous to existing RVOS works, the pipeline consists of three main procedures: feature encoding, multimodal feature interaction, and feature decoding. Albeit various strategies have been implemented to improve the interaction, they fail to utilise multimodal features comprehensively due to the involvement of intertwined textual components.

To mitigate this issue, we propose to decouple the subject and context during the interaction. Specifically, we first encode visual and text features individually. Then, a subject perceptron is formed to separate the subject from the encoded text features. Next, the subject features and sentence features are fed into the decoupled multimodal interaction module, which consists of two parallel branches: (1) Subject-aware interaction and (2) Context-aware interaction. Each branch sequentially performs inter-modal feature fusion and intra-modal feature refinement. Finally, we concatenate the outputs of both branches and utilise them to generate masks for the text-referred objects.

B. Subject Perceptron

From the pipeline, it is evident that subject extraction is crucial to our RVOS architecture. In this paper, we achieve this via a subject perceptron, based on self-attention over token-level text features. As shown in Fig. 3 (a), given the encoded text features $\{e_l\}_{l=1}^{L}, e_l \in \mathbb{R}^C$, the subject perceptron assigns each token a probability $s_l \in \mathbb{R}^2$ that belongs to the subject of the text:

$$\begin{cases} \{e'_l\}_{l=1}^{L} = LN(MHSA(\{e_l\}_{l=1}^{L}) + \{e_l\}_{l=1}^{L}), \\ \{s_l\}_{l=1}^{L} = Pred(LN(FFN(\{e'_l\}_{l=1}^{L}) + \{e'_l\}_{l=1}^{L})), \end{cases} \tag{1}$$

where L and C are the number of word tokens and feature channels. Note that L usually does not reflect the length of the input sentence since some words would be parsed into several tokens during tokenisation. Specifically, given an input text, we first utilise a tokeniser to convert it to a sequence of smaller semantic units (tokens), which are then fed into the text encoder for $\{e_l\}_{l=1}^{L}, e_l \in \mathbb{R}^C$. $MHSA$, LN, FFN, and $Pred$ define the modules for multi-head self-attention, linear normalisation, feed-forward network, and prediction, respectively. To facilitate probability generation, the prediction module is built by concatenating a linear layer and a softmax layer. During training, the above modules are optimised by minimising a cross entropy loss:

$$\mathcal{L}_{\text{Sub}} = \frac{1}{L}\Sigma_{l=1}^{L}\mathcal{L}_{\text{CrossEntropy}}(s_l, s_{\text{GT},l}), \tag{2}$$

where $s_{\text{GT},l} \in \{0, 1\}$ is the ground truth label for the l^{th} word token. The positive labels indicate the location of the subject part in the text. Since the subject information is not provided in current RVOS datasets, we utilise the existing annotations (pixel-level masks-sentence pairs) and a pre-trained vision-language model (GLIP [24]) to generate pseudo labels.

As shown in Fig. 3 (b), given a sentence and a video frame with pixel-level masks, we leverage the powerful alignment between object-level and phrase-level features of GLIP [24] to extract the subject part from the sentence. At first, we generate the object region (red bounding box) from the mask and divide the input sentence into noun phrases (via Spacy). Then, we embed object-level features $o \in \mathbb{R}^C$ and phrase-level features $\{n_l\}_{l=0}^{L_n} \in \mathbb{R}^C$ with GLIP vision and text

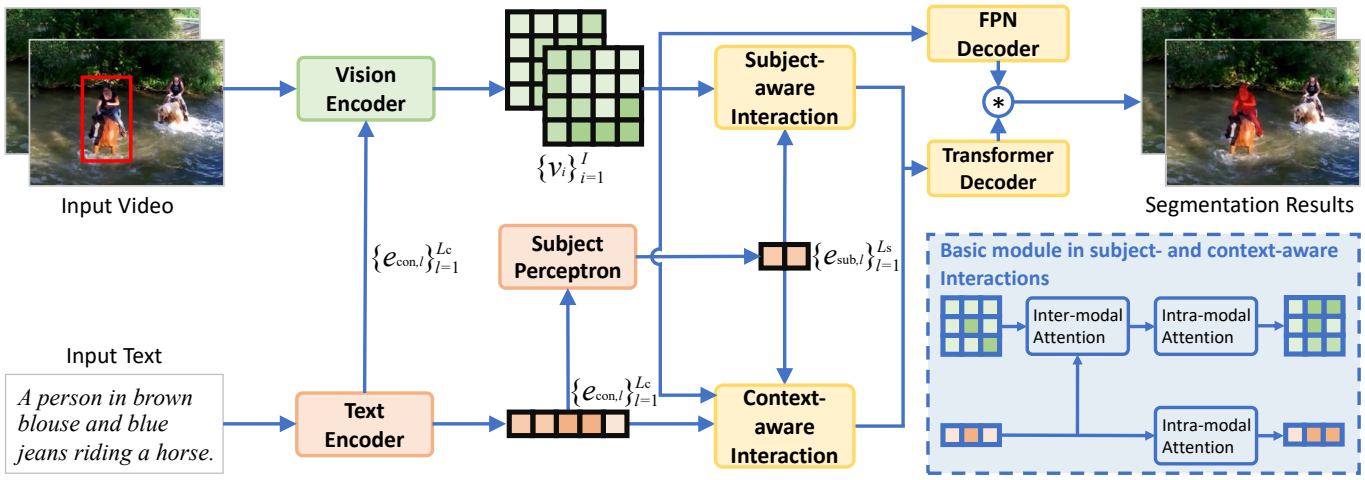

Fig. 2. Pipeline of our proposed Decoupled Multimodal Transformers (DMFormer) for RVOS. The red box highlights the target object referred to by the input text. The blue box with dotted boundary details the basic module in subject-aware and context-aware interactions. * refers to the convolution operation.

encoders, respectively. C and L_n denote the number of feature channels and the parsed noun phrases. Next, we compute the similarities between o and $\{n_l\}_{l=0}^{L_n}$, achieving $\{p_l\}_{l=0}^{L_n}$. With the powerful capabilities for feature alignment, high-quality p_l can be achieved via GLIP. This way, the subject part can be extracted by selecting the phrase with the highest similarity. Finally, we generate the pseudo labels $s_{\text{GT},l}$:

$$s_{\text{GT},l} = \begin{cases} 0, & l \notin \mathcal{P}, \\ 1, & l \in \mathcal{P}, \end{cases} \tag{3}$$

where $\mathcal{P}$ refers to the token set of the phrase that aligns the best with the target object. Note that we only employ Equations 2 and 3 and GLIP vision and text encoders during training. For each iteration, we perform referring segmentation on a synthesised video, which consists of 3 frames sampled from image/video datasets. To generate accurate pseudo labels, we consider the correspondence between noun phrases and object regions from all 3 video frames. For inference, we directly predict s_l from the input sentence and then select the segment corresponding to the highest probability.

C. Decoupled Multimodal Interaction

With the subject perception, our method can selectively align visual features with different textual components, achieving the decoupled multimodal interaction. This is under-explored in current RVOS works. As shown in Fig. 2, our decoupled architecture mainly consists of two parallel branches: subject-aware interaction and context-aware interaction. Each branch has two concatenated basic modules. The main difference between the two branches lies in the text input. The subject-aware interaction takes subject features $\{e_{\text{sub},l}\}_{l=1}^{L_s}$ to interact with visual features. By contrast, the context-aware interaction considers all text features (or context features, $\{e_{\text{con},l}\}_{l=1}^{L_c}$), leveraging the relationship between the subject and other components during the interaction. L_s and L_c are the number of subject and all word tokens.

Here we introduce the feed-forward procedure of one basic interaction module, which is highlighted with the blue box in Fig. 2. Given the text features $\{e_l\}_{l=1}^{L}, e_l \in \mathbb{R}^C$ (e_l could be the subject or context features) and a set of multi-scale visual features $\{v_i\}_{i=1}^{I}, v_i \in \mathbb{R}^{T\times H_i\times W_i\times C}$ (encoded from the input video sequence including T frames), the module first performs inter-modal feature fusion:

$$v_i = LN(MHCA(v_i, \{e_l\}_{l=1}^{L}) + v_i), \tag{4}$$

where I and C are the number of spatial scales and feature channels. $H_i \times W_i$ represents the spatial dimension on the i^{th} scale. $MHCA$ denotes multi-head cross attention, where keys and values come from textual features $\{e_l\}_{l=1}^{L}$. They are queried by different scales of visual features v_i. After this, LN is performed for linear normalisation. During this stage, visual features are aggregated under the guidance from the subject and context components from the text, respectively. Since the aggregation is performed on spatial-temporal visual features, both static and motion clues relevant to the text can be mined implicitly, facilitating the localisation of the target object in the input video sequence.

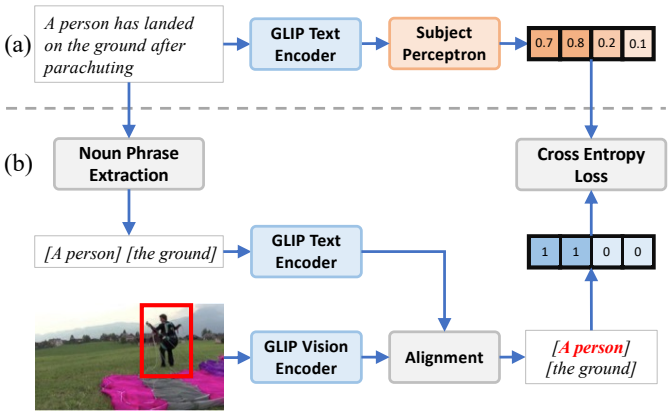

Fig. 3. Illustration of the subject perceptron. (a) A feed-forward example of the subject perceptron; (b) Pseudo label generation from the given text and ground truth object box. Note that the probabilities of all text components are individual and we only generate pseudo labels in the training stage.

After feature aggregation, $\{v_i\}_{i=1}^{I}$ and $\{e_l\}_{l=1}^{L}$ are fed into the self-attention modules to mine their intra-modal properties. Specifically, we handle visual and text features using the same attention mechanism as Equation 1 (besides the $Pred$ module). The self-attention results are the outputs of one basic interaction module, which have the same dimensions as the input visual and text features. Therefore, multiple modules could be concatenated for decoupled and deep multimodal interactions. The main goal of alternate intra-modal and inter-modal modules is to interact between multimodal features while simultaneously keeping their intra-modal properties. Both the subject-aware and context-aware interaction branches in our DMFormer consist of two basic modules for a better trade-off between accuracy and efficiency.

With the output features from both subject- and context-aware interaction branches: $\{v_{\text{sub},i}\}_{i=1}^{I} \in \mathbb{R}^{T\times H_i\times W_i\times C}$, $\{e_{\text{sub},l}\}_{l=1}^{L_s} \in \mathbb{R}^{L_s\times C}$, $\{v_{\text{con},i}\}_{i=1}^{I} \in \mathbb{R}^{T\times H_i\times W_i\times C}$, and $\{e_{\text{con},l}\}_{l=1}^{L_c} \in \mathbb{R}^{L_c\times C}$, we have achieved the decoupling of multimodal interactions. After this, we decode the outputs to generate the spatial-temporal object masks. At first, we form a post-processor to resize and concatenate the outputs:

$$\begin{cases} v_{\text{dec},i} = O(D(v_{\text{sub},i}), D(v_{\text{con},i})), \\ e_{\text{dec}} = O(P(D(\{e_{\text{sub},l}\}_{l=1}^{L_\text{s}})), P(D(\{e_{\text{con},l}\}_{l=1}^{L_\text{c}}))), \end{cases} \tag{5}$$

where D, P, and O define resizing, pooling, and concatenation operations, respectively. D is implemented by an MLP module, which down-samples feature channels by half. P denotes the average pooling and operates on text features only. Given the resized token-level subject features $D(\{e_{\text{sub},l}\}_{l=1}^{L_\text{s}}) \in \mathbb{R}^{L_\text{s}\times C/2}$ or context features $D(\{e_{\text{con},l}\}_{l=1}^{L_\text{c}}) \in \mathbb{R}^{L_\text{c}\times C/2}$, P performs average pooling along the token dimension, achieving a single vector with the dimension of $C/2$. O concatenates subject and context features along channels. After this, the decoupled features can be achieved, achieving $v_{\text{dec},i} \in \mathbb{R}^{T\times H_i\times W_i\times C}$ and $e_{\text{dec}} \in \mathbb{R}^{C}$, to facilitate the following mask generation. Note that albeit the outputs from different branches are integrated, the decoupled properties are not disturbed so they still benefit the overall architecture.

Upon feature integration, we generate object masks for the text-referred object throughout the video. Analogous to the previous works based on DETR architectures [17]–[19], we build our decoder with the deformable decoder and dynamic convolution. During inference, we first feed the integrated visual features $\{v_{\text{dec},i}\}_{i=1}^{I}$ and text features e_{dec} into the deformable decoder, where the former conveys spatial-temporal clues of the input video. The latter is repeated and concatenated with fixed-number of learnable vectors and serves as queries to generate object-level outputs across frames. Next, the outputs are further decoded with the object head, box head, and mask head, respectively. The first two measures the probabilities that belong to the target and object locations. The mask head predicts parameters for a set of kernels, which convolve multi-scale visual features to generate pixel-level masks. After filtering with the probabilities from the object head, the final predictions can be achieved as the RVOS results.

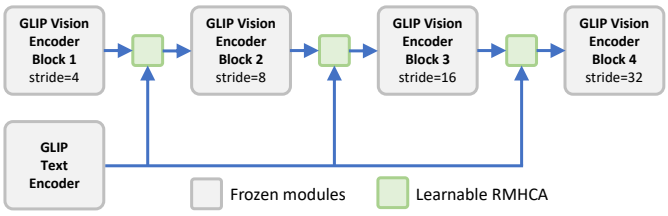

Fig. 4. Illustration of VLP knowledge incorporation, based on the vision and text encoders from GLIP [24]. To better interact multimodal features, we perform RMHCA after each block since they have different output dimensions (1/4, 1/8, 1/16, and 1/32 of the resolutions of the input frame), implicitly encoding different semantic levels. More details about the learnable RMHCAs are formulated in Equation 6.

D. Incorporation of VLP knowledge

To better align visual and text features during the interaction, we leverage a large-scale pre-trained vision-language model (GLIP [24]) to implement and initialise our visual and text encoders. Despite GLIP and our architecture considering different tasks (phrase grounding and referring segmentation), the proposed decoupling can mitigate such a gap. This is because, in the training stage, the decoupled multimodal interaction implicitly decomposes RVOS into two separate tasks: subject phrase grounding and discriminative feature embedding. Such a decomposition enables the targets of GLIP and ours closer and thus facilitates the knowledge transfer between them. To improve the transfer, we insert the residual multi-head cross-attention (termed as RMHCA) layers into the late stages of the visual encoder, further aligning visual and text features and facilitating the interaction between them. The main idea of RMHCA is shown in Fig. 4.

Given the intermediate visual features $v_i \in \mathbb{R}^{T\times H_i\times W_i\times C_i}$ and the encoded text features $\{e_l\}_{l=1}^{L}, e_l \in \mathbb{R}^{C}$, our RMHCA updates v_i as follows:

$$v_i = L_{\text{out,i}}(LN(RMHCA(L_{\text{in,i}}(v_i), \{e_l\}_{l=1}^{L}) + L_{\text{in,i}}(v_i))), \tag{6}$$

where i indicates different feature levels. T, H_i, W_i, and C_i denote the temporal, height, width, and channel dimensions on the i^{th} level. LN is the linear normalisation module. To align channel dimensions between visual and textual features, we employ two linear layers: $L_{\text{in,i}}$ and $L_{\text{out,i}}$, which map the number of channels of v_i from C_i to C, and from C to C_i for seamless integration of RMHCA.

E. Implementation Details

The proposed DMFormer consists of four trainable modules: (1) residual multi-head cross-attention (RMHCA), (2) subject perceptron, (3) decoupled multimodal interaction, and (4) decoder. We fix visual and text encoders to maintain the transferred knowledge. During training, all these modules are trained end-to-end, by minimising the following loss:

$$\mathcal{L} = \lambda_{\text{Sub}}\mathcal{L}_{\text{Sub}} + \lambda_{\text{Cls}}\mathcal{L}_{\text{Cls}} + \lambda_{\text{Box}}\mathcal{L}_{\text{Box}} + \lambda_{\text{Mask}}\mathcal{L}_{\text{Mask}}. \tag{7}$$

Following previous works [17], [18], we compute $\mathcal{L}_{\text{Cls}}$, $\mathcal{L}_{\text{Box}}$, and $\mathcal{L}_{\text{Mask}}$ by evaluating the pixel-level, box-level, and class-level similarities between the predictions and ground

truth. $\mathcal{L}_{\text{Sub}}$ evaluates the predictions from the subject perceptron. More details can be found in Equation 2. Besides the outputs from the last decoder layer, the intermediate decoding results are also considered. Specifically, these results are also decoded to pixel-level masks and compared with the ground truth, using $\mathcal{L}_{\text{Cls}}$, $\mathcal{L}_{\text{Box}}$, and $\mathcal{L}_{\text{Mask}}$. This way, more consistent predictions can be achieved for the target object. In this paper, we set λ_{Sub}, λ_{Cls}, λ_{Box} and λ_{Mask} as 2, 2, 5, and 2.

For the overall architecture, we consider Swin-Transformer (Tiny/Large) [35] and BERT (Base) [36] as visual and text encoders, which come from the GLIP model [24] and remain frozen during training and inference to keep the powerful capabilities for vision-language alignment. All multi-head attention modules have 8 heads and the input/output feature dimensions are 256. We utilise four deformable decoders to form the multimodal decoder. We implement DMFormer with PyTorch [37] and perform training and evaluation on Nvidia RTX A6000 GPUs. The inference details of our DMFormer is illustrated in Algorithm 1.

The model is first pre-trained on image datasets for referring segmentation, including RefCOCO [38], RefCOCO+ [38], and RefCOCOg [39]. Then, following previous works, we fine-tune the model on A2D Sentence [3] for the evaluation on A2D Sentence [3] and J-HMDB Sentence [3]. When testing on Ref-YouTube-VOS [21] and Ref-DAVIS-17 [20], the model is fine-tuned on Ref-YouTube-VOS [21]. During pre-training, we optimise the model for 10 epochs, with a learning rate of 1e-4 (multiplied by 0.1 from the 6^{th} and 8^{th} epoch). During fine-tuning, we optimise the model for 6 epochs, with a learning rate of 1e-4 (multiplied by 0.1 from the 3^{rd} and 5^{th} epoch). As for the subject perceptron, it is learnable during both pre-training and fine-tuning. To generate high-quality pseudo labels, we employ GLIP [24] with the largest scale of pre-trained architecture (with Swin-Transformer-Large and BERT-base) to embed and align visual and text features.

Algorithm 1 Decoupled Multimodal Transformer for RVOS.

INPUT: A video sequence $\mathcal{V}$ and a language expression $\mathcal{E}$ referring the target object.
OUTPUT: Target object masks $\mathcal{M}$ on all video frames.

Vision and text feature embedding
$\{e_l\}_{l=1}^{L} \leftarrow \texttt{TextBackbone}(\mathcal{E})$
$\{v_i\}_{i=1}^{I} \leftarrow \texttt{VisionBackbone}(\{e_l\}_{l=1}^{L}, \mathcal{V})$
Subject perception and separation
$\{p_n\}_{n=1}^{N} \leftarrow \texttt{FindNounPhrase}(\mathcal{E})$
$\{s_l\}_{l=1}^{L} \leftarrow \texttt{SubjectPerceptron}(\{e_l\}_{l=1}^{L})$
$\{e_{\text{sub},l}\}_{l=1}^{L_{\text{s}}} \leftarrow \texttt{Separation}(\{e_l\}_{l=1}^{L}, \{p_n\}_{n=1}^{N}, \{s_l\}_{l=1}^{L})$
$\{e_{\text{con},l}\}_{l=1}^{L_{\text{c}}} \leftarrow \{e_l\}_{l=1}^{L}$
Subject-aware Interaction
$\{v_{\text{sub},i}\}_{i=1}^{I} \leftarrow \{v_i\}_{i=1}^{I}$
for $itr \leftarrow 1$ **to** 2 **do**
  $\{v_{\text{sub},i}\}_{i=1}^{I} \leftarrow \texttt{MHCA}(\{v_{\text{sub},i}\}_{i=1}^{I}, \{e_{\text{sub},l}\}_{l=1}^{L_{\text{s}}})$
  $\{v_{\text{sub},i}\}_{i=1}^{I} \leftarrow \texttt{MHSA}(\{v_{\text{sub},i}\}_{i=1}^{I})$
  $\{e_{\text{sub},l}\}_{l=1}^{L_{\text{s}}} \leftarrow \texttt{MHSA}(\{e_{\text{sub},l}\}_{l=1}^{L_{\text{s}}})$
Context-aware Interaction
$\{v_{\text{con},i}\}_{i=1}^{I} \leftarrow \{v_i\}_{i=1}^{I}$
for $itr \leftarrow 1$ **to** 2 **do**
  $\{v_{\text{con},i}\}_{i=1}^{I} \leftarrow \texttt{MHCA}(\{v_{\text{con},i}\}_{i=1}^{I}, \{e_{\text{con},l}\}_{l=1}^{L_{\text{c}}})$
  $\{v_{\text{con},i}\}_{i=1}^{I} \leftarrow \texttt{MHSA}(\{v_{\text{con},i}\}_{i=1}^{I})$
  $\{e_{\text{con},l}\}_{l=1}^{L_{\text{s}}} \leftarrow \texttt{MHSA}(\{e_{\text{con},l}\}_{l=1}^{L_{\text{c}}})$
Feature integration
$\{v_{\text{dec},i}\}_{i=1}^{I} \leftarrow$
  $[\texttt{Resize}(\{v_{\text{sub},i}\}_{i=1}^{I}), \texttt{Resize}(\{v_{\text{con},i}\}_{i=1}^{I})]$
$e_{\text{dec}} \leftarrow [\texttt{AvgPool}(\texttt{Resize}(\{e_{\text{sub},l}\}_{l=1}^{L_{\text{s}}})),$
  $\texttt{AvgPool}(\texttt{Resize}(\{e_{\text{con},l}\}_{l=1}^{L_{\text{c}}}))]$
Mask Generation
$\mathcal{K} \leftarrow \texttt{DETRDecoder}\{v_{\text{dec},i}\}_{i=1}^{I}, e_{\text{dec}})$
$\{v_{\text{fpn},i}\}_{i=1}^{I} \leftarrow$
  $\texttt{CrossModalFPN}\{v_{\text{dec},i}\}_{i=1}^{I}, \{v_i\}_{i=1}^{I}, \{e_l\}_{l=1}^{L})$
$\mathcal{M} \leftarrow \texttt{Sigmoid}(\texttt{Convolution}(\mathcal{K}, \{v_{\text{fpn},i}\}_{i=1}^{I}))$

IV. Experiments

A. Datasets and Metrics

1) Datasets: We evaluate our method on four benchmark datasets for RVOS: A2D Sentences [3], J-HMDB Sentences [3], Ref-DAVIS-17 [20], and Ref-YouTube-VOS [21], which respectively contain 3782, 928, 90, and 4519 videos, with 6656, 928, 1544, and 12913 language expressions. Each expression refers to one target object. Among the datasets, A2D Sentences and J-HMDB Sentences are characterised by action-orient descriptions since their videos are collected from the datasets for actor and action segmentation. By contrast, Ref-DAVIS-17 and Ref-YouTube-VOS describe target objects with unconstrained words. As for the expression length, Ref-YouTube-VOS has more words on average (9.75) than others (A2D Sentences: 6.94, J-HMDB Sentences: 6.15, Ref-DAVIS-17: 7.03), making it more challenging. Therefore, the comparison results and ablations in this section are mainly achieved on Ref-YouTube-VOS to illustrate better the performance improvement brought by the proposed idea.

2) Evaluation metrics: Following current RVOS works, we apply different evaluation metrics on different benchmarks. For A2D Sentences and J-HMDB Sentences, three metrics are considered: Precision@K, mAP (mean Average Precision), overall IoU (Intersection over Union), and mean IoU. Specifically, Precision@K measures the percentage of test samples whose overlap values are higher than the threshold K (0.5:0.1:0.9). mAP is measured over 0.50:0.05:0.90. The overall IoU is measured over all test samples and mean IoU averages the IoU of each sample. When evaluating on Ref-DAVIS-17 and Ref-YouTube-VOS, we employ Jaccard-Index ($\mathcal{J}$) and F-measure ($\mathcal{F}$) to evaluate the segmentation quality in IoU and boundary, respectively. Their mean $\mathcal{J}\&\mathcal{F}$ denotes the overall performance.

B. Comparison with State-of-the-art Methods

1) Ref-YouTube-VOS & Ref-DAVIS-17: We first compare our proposed DMFormer with the state-of-the-arts on Ref-YouTube-VOS [21] and Ref-DAVIS-17 [20]. The quantitative results are shown in Table I. Since GLIP [24] only provides the knowledge on Swin Transformer-Tiny/Large [35], we

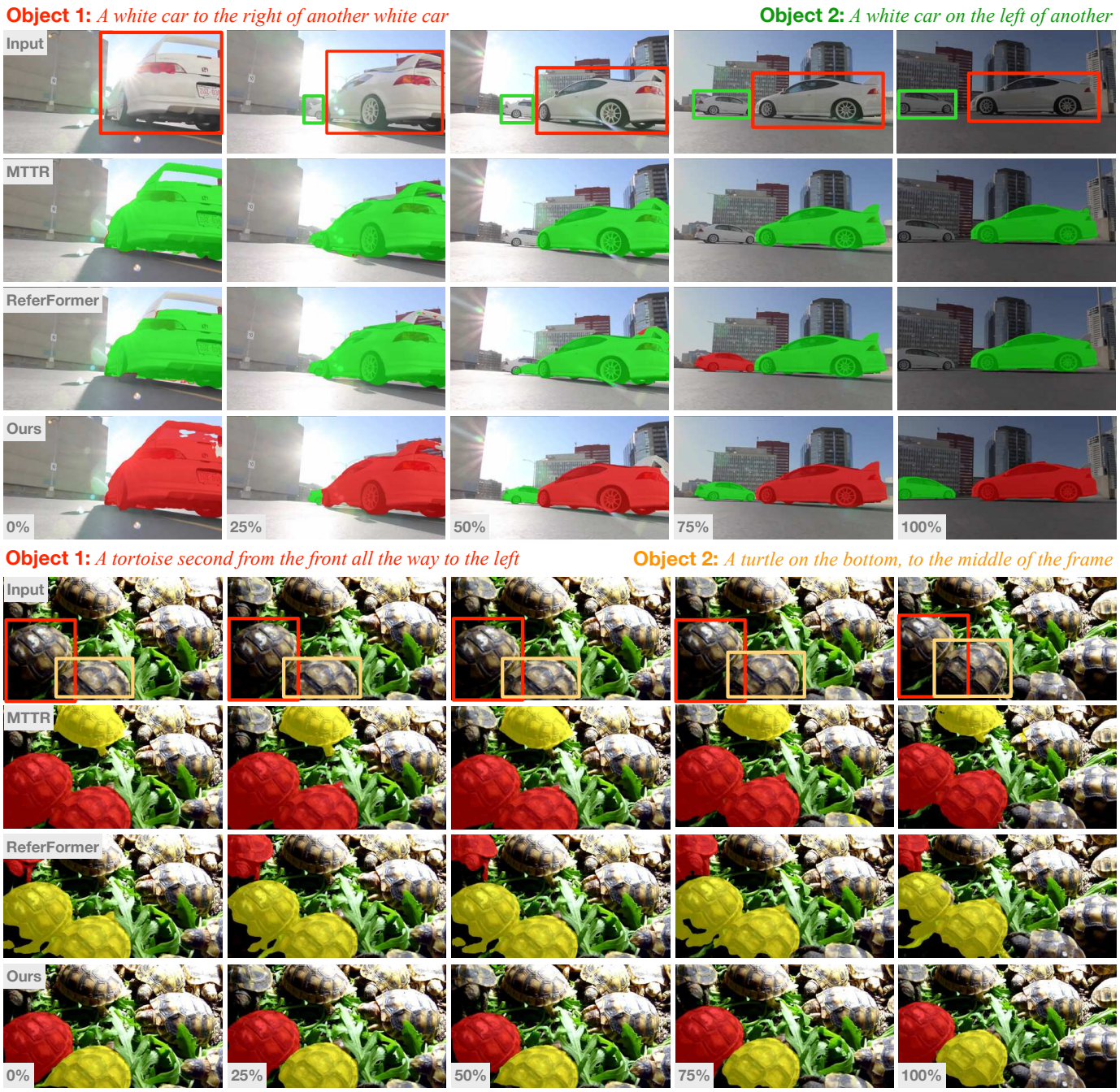

Fig. 5. Qualitative results on two video sequences from Ref-YouTube-VOS [21]. For each sequence, we show the input frames in the top row, with the bounding boxes of the text-referred (target) objects. The percentage indicates the position of each frame in the sequence.

report the RVOS performance with these backbones. To better validate our performance improvement, we focus on the results of the previous SoTA model (ReferFormer [18]), which employs the same backbones as ours. With respective backbones, DMFormer outperforms ReferFormer on both Ref-YouTube-VOS and Ref-DAVIS-17 in all evaluation metrics, validating its effectiveness and consistent improvement. In addition, although MTTR [17] is built on a larger visual backbone (Video Swin-Transformer-Tiny), the proposed DMFormer with Swin-Tranformer-Tiny achieves better performance, which implicitly demonstrates that DMFormer can learn spatial-temporal clues from input videos and interact well with input texts.

Fig. 5 and Fig. 6 compare the qualitative results from our DMFormer (Swin-Ttransformer-Tiny), ReferFormer [18] (Swin-Ttransformer-Tiny), and MTTR [17] (Video-Swin-Ttransformer-Tiny) on Ref-YouTube-VOS [21]. As shown in Fig. 5, our model performs more robustly against the videos with multiple distractors, showing the proposed decoupled interaction can better leverage text clues than previous works. With the VLP knowledge, our model can perceive more vision-language alignments to facilitate RVOS. As a result, better performance can be achieved on videos with complex scenes

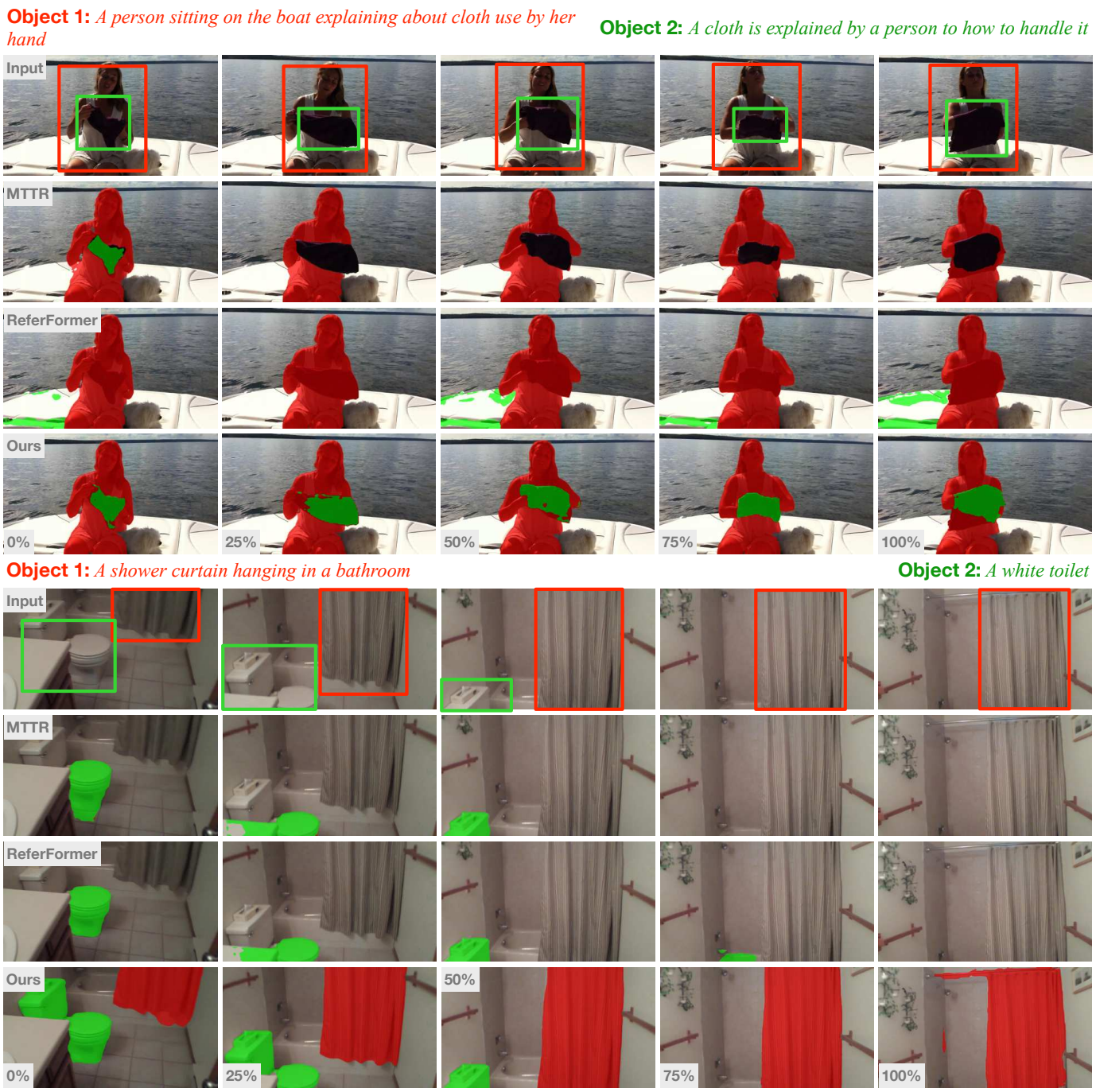

Fig. 6. Qualitative results on two video sequences from Ref-YouTube-VOS. For each sequence, we show the input frames in the top row, with the bounding boxes of the text-referred (target) objects. The percentage indicates the position of each frame in the sequence.

and language expressions, as shown in the first row in Fig. 6. In addition, the VLP knowledge brings excellent performance on unseen/rare objects in the RVOS training data. For example, the "curtain" in the bottom video sequence in Fig. 6, where both MTTR and ReferFormer entirely ignore the curtain area since the relevant vision-language alignments rarely appear in the training data.

2) A2D Sentences & J-HMDB Sentences: Tables II and III show the comparison results on A2D Sentences [3] and J-HMDB Sentences [3], respectively. It is observed that the performance gaps between the state-of-the-arts and ours are marginal. This is mainly because the involved datasets are less challenging than Ref-YouTube-VOS and Ref-DAVIS-17. Still, DMFormer outperforms others on almost all evaluation metrics (besides the overall IoU on A2D Sentences). This shows that ReferFormer performs favourably on several big objects but not most ones since DMFormer outperforms ReferFormer in the mean IoU. Therefore, the performance improvement brought by our proposed idea is consistent.

TABLE I
QUANTITATIVE COMPARISON OF DIFFERENT METHODS ON REF-YOUTUBE-VOS [21] AND REF-DAVIS-17 [20]. THE TOP 2 SCORES ARE HIGHLIGHTED IN RED AND BLUE.

Method	Backbone	Ref-YouTube-VOS			Ref-DAVIS-17		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
URVOS [21]	ResNet-50	47.2	45.3	49.2	–	–	–
CMPC-V [7]	I3D	47.5	45.6	49.3	–	–	–
YOFO [10]	ResNet-50	48.6	47.5	49.7	53.3	48.8	57.8
LBDT [12]	ResNet-50	49.4	48.1	50.6	54.5	–	–
LOCATOR [16]	Transformers	50.0	48.8	51.1	–	–	–
MTTR [17]	V-Swin-T	55.3	54.0	56.6	–	–	–
ReferFormer [18]	Swin-T	58.7	57.6	59.9	55.8	53.2	58.3
ReferFormer [18]	Swin-L	62.4	60.8	64.1	60.5	57.6	63.4
VLT [15]	V-Swin-B	63.8	61.9	65.6	61.6	58.9	64.3
DMFormer	Swin-T	61.4	60.1	62.7	57.4	54.9	59.9
DMFormer	Swin-L	64.9	63.4	66.5	62.3	59.5	65.1

TABLE II
QUANTITATIVE COMPARISON OF DIFFERENT METHODS ON A2D SENTENCES [3]. THE TOP 2 SCORES ARE HIGHLIGHTED IN RED AND BLUE.

Method	Backbone	Precision					IoU		mAP
		P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	Overall	Mean	0.5:0.95
CMSA-V [8]	ResNet-101	48.7	43.1	35.8	23.1	5.2	61.8	43.2	–
CMPC-V [7]	I3D	65.5	59.2	50.6	34.2	9.8	65.3	57.3	40.4
CSTM [9]	I3D	65.4	58.9	49.7	33.3	9.1	66.2	56.1	39.9
LBDT [12]	ResNet-50	73.0	67.4	59.0	42.1	13.2	70.4	62.1	47.2
MMVT [11]	ResNet-101	64.5	59.7	52.3	37.5	13.0	67.3	55.8	41.9
LOCATOR [16]	Transformers	70.9	64.0	52.5	35.1	10.1	69.0	59.7	46.5
MTTR [17]	V-Swin-T	75.4	71.2	63.8	48.5	16.9	72.0	64.0	46.1
ReferFormer [18]	Swin-T	80.9	77.6	70.7	54.1	19.4	75.9	68.0	52.7
ReferFormer [18]	Swin-L	83.5	80.6	74.3	58.2	22.0	78.8	70.5	55.4
DMFormer	Swin-T	81.3	78.8	71.9	55.2	20.3	76.0	68.3	54.3
DMFormer	Swin-L	83.7	81.8	75.7	60.0	24.3	78.4	70.9	58.2

C. Ablation Studies

To validate the contributions of the proposed idea on the overall architecture, we conduct a series of ablation studies. At first, we analyse the role of the subject perceptron. Then, the effectiveness of the decoupled multimodal interaction is given via both quantitative and qualitative results. Finally, we illustrate the performance gain brought by VLP incorporation. In this section, we consider Swin-Transformer-Tiny [35] and BERT-Base [36] as vision and language backbones, respectively. All models are first pre-trained on RefCOCO [38], RefCOCO+ [38], and RefCOCOg [39], and then fine-tuned on Ref-YouTube-VOS [21]. The experimental results are evaluated on Ref-YouTube-VOS [21].

1) Subject Perceptron: This module is critical to the overall architecture as it determines the way we decouple multimodal interactions. In this paper, we propose a self-attention-based module to locate the subject part from the input text. The module is optimised with other modules in DMFormer, achieving end-to-end training. To validate the effectiveness of the subject perceptron, we keep other modules in DMFormer unchanged and implement a non-trainable method (based on the NLTK tookit [40]) to perform subject perception. The comparison results is shown in Table IV. From the table, the model in the first row is ReferFormer [18], which serves as the baseline. The second model employs the decoupled multimodal interaction but no subject perceptron. In this case, both subject- and context-aware interactions take all word tokens. Despite this, better results can be achieved than the baseline, showing the excellent potential of the decoupled interaction. When equipped with the non-trainable tool, we found the overall performance drops significantly due to the texts with diverse structures. This illustrates that the proposed subject perceptron can handle better the input texts and adapt well with the decoupled interaction.

2) Doupled Multimodal Interaction: Next, we show the effectiveness of the decoupled multimodal interaction. As shown in Table IV, the decoupled interaction brings better results even without a subject perceptron. Despite taking the same textual inputs, we conjuncture that the decoupled structure can still be trained to focus on different and complementary multimodal interactions. With the well-trained subject perceptron, the RVOS performance is further improved, which shows that the syntactic structure of the text indeed benefits the decoupled multimodal interaction.

To demonstrate the impact of the decoupled multimodal interaction more intuitively, we select some hard samples with multiple distractors from Ref-YouTube-VOS [21], where we perform RVOS using the DMFormer with/without the decoupled interaction (both with the VLP initialised knowledge). The qualitative comparison results between the variants are shown in Fig. 7. From the figure, it is observed

TABLE III
QUANTITATIVE COMPARISON OF DIFFERENT METHODS ON J-HMDB SENTENCES [3]. THE TOP 2 SCORES ARE HIGHLIGHTED IN RED AND BLUE.

Method	Backbone	Precision					IoU		mAP
		P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	Overall	Mean	0.5:0.95
CMSA-V [8]	ResNet-101	76.4	62.5	38.9	9.0	0.1	62.8	58.1	–
CMPC-V [7]	I3D	83.1	65.7	37.1	7.0	0.0	61.6	61.7	34.2
CSTM [9]	I3D	78.3	63.9	37.8	7.6	0.0	59.8	60.4	33.5
LBDT [12]	ResNet-50	86.4	74.4	53.3	13.2	0.0	64.5	65.8	41.1
MMVT [11]	ResNet-101	79.9	71.4	49.0	12.6	0.1	61.9	61.3	38.6
LOCATOR [16]	Transformers	89.3	77.2	50.8	10.6	0.2	67.3	66.3	45.6
MTTR [17]	V-Swin-T	93.9	85.2	61.6	16.6	0.1	70.1	69.8	39.2
ReferFormer [18]	Swin-T	94.7	87.3	65.2	18.6	0.3	71.6	70.4	41.5
ReferFormer [18]	Swin-L	96.9	90.0	70.1	21.7	0.3	73.3	72.1	42.7
DMFormer	Swin-T	95.2	88.5	66.4	20.1	0.3	71.9	70.5	42.5
DMFormer	Swin-L	97.2	92.5	72.1	23.4	0.3	73.9	72.8	44.7

TABLE IV
THE EFFECTIVENESS OF THE SUBJECT PERCEPTRON AND THE DECOUPLED MULTIMODAL INTERACTION. THE FIRST AND SECOND COLUMNS INDICATE THE USAGE OF THE SUBJECT PERCEPTRON AND DECOUPLED MULTIMODAL INTERACTION, RESPECTIVELY.

Subject perceptron	Decouple	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
✗	✗	58.7	57.4	60.1
✗	✓	59.1	57.7	60.5
NLTK tools	✓	54.0	51.1	54.9
Ours	✓	59.7	58.5	60.9

TABLE V
THE EFFECTIVENESS OF VLP INCORPORATION. THE FIRST, SECOND, AND THIRD COLUMNS INDICATE THE USAGE OF THE VLP KNOWLEDGE, RESIDUAL MULTI-HEAD CROSS-ATTENTION, AND DECOUPLED MULTIMODAL INTERACTION, RESPECTIVELY.

VLP	RMHCA	Decouple	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
✗	✗	✗	58.7	57.4	60.1
✓	✗	✗	60.5	59.2	61.7
✓	✓	✗	60.9	59.7	62.1
✓	✓	✓	61.4	60.1	62.7

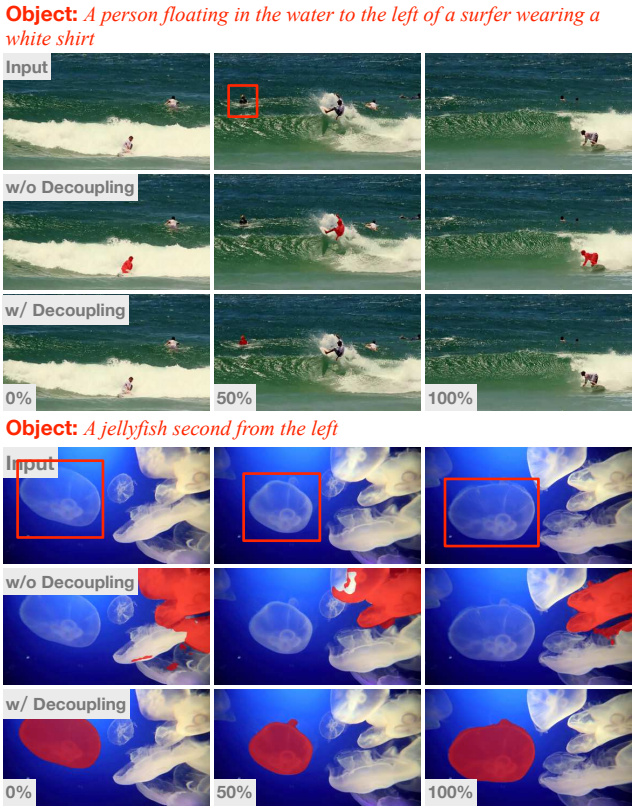

Fig. 7. Visualised ablations for the decoupled multimodal interaction. For each sample, the top row shows the input video sequences and highlights the text-referred (target) objects (red boxes). Note the frame without boxes means the target object does not appear. The red masks in the middle and bottom rows are the segmentation results. The percentage indicates the position of each frame in the video.

that the variant without the decoupled interaction struggles with complex language expressions and distractors. This is mainly because the model focuses more on the subject and ignores other descriptions, which are crucial for suppressing distractors, especially the ones with the same categories. In contrast, with the decoupled subject-aware and context-aware multimodal interactions, DMFormer fairly considers different textual components. This way, more focus can be made on the discriminative information, which facilitates the robust RVOS on these challenging samples.

3) VLP knowledge: Finally, we illustrate the performance improvement brought by the large-scale pre-trained vision-language alignment [24]. More details are shown in Table V, where the first row is the baseline (ReferFormer [18]). From the second row, it is observed that even with the task gap between referring segmentation and phrase grounding (where the VLP knowledge are pre-trained), the large-scale pre-trained vision-language alignment can improve the RVOS performance of the baseline. This illustrates the vast potential of the application of VLP knowledge to RVOS. The performance can be improved by incorporating the residual multi-head attention (RMHCA) modules in the encoding stage, which further solidates the alignment between vision and language modalities. The results in the bottom row show that the RVOS performance can be further boosted, under the decoupled multi-modal interaction. This shows that the decoupled interaction can implicitly shrink the task gap so that the large-scale pre-trained vision-language alignment can be better incorporated into the RVOS architecture.

V. Conclusion

In this paper, we presented the decoupled multimodal transFormer (DMFormer) for RVOS. Given the input video sequence and text, we first perceived the subject part from the text, via a learnable subject perceptron. Then, the separated text features were fed into two parallel branches for subject- and context-aware multi-modal interactions. Finally, we decoded the interaction results for the final predictions. With the decoupled interactions, our RVOS architecture is encouraged to focus on more comprehensive relationships among features. In addition, it also facilitates the incorporation of large-scale pre-trained vision-language alignment. Experimental results show that the proposed method outperforms the state-of-the-art on all RVOS benchmarks. We hope that the proposed decoupling idea and VLP knowledge incorporation can inspire future contributions.

References

[1] H.-C. Shih, "A survey of content-aware video analysis for sports," *IEEE TCSVT*, vol. 28, no. 5, pp. 1212–1231, 2017.

[2] W. Wang, T. Zhou, F. Porikli, D. Crandall, and L. Van Gool, "A survey on deep learning technique for video segmentation," *arXiv preprint arXiv:2107.01153*, 2021.

[3] K. Gavrilyuk, A. Ghodrati, Z. Li, and C. G. Snoek, "Actor and action video segmentation from a sentence," in *CVPR*, 2018, pp. 5958–5966.

[4] H. Wang, C. Deng, F. Ma, and Y. Yang, "Context modulated dynamic networks for actor and action video segmentation with language queries," in *AAAI*, vol. 34, 2020, pp. 12 152–12 159.

[5] H. Wang, C. Deng, J. Yan, and D. Tao, "Asymmetric cross-guided attention network for actor and action video segmentation from natural language query," in *ICCV*, 2019, pp. 3939–3948.

[6] K. Ning, L. Xie, F. Wu, and Q. Tian, "Polar relative positional encoding for video-language segmentation," in *IJCAI*, 2021, pp. 948–954.

[7] S. Liu, T. Hui, S. Huang, Y. Wei, B. Li, and G. Li, "Cross-modal progressive comprehension for referring segmentation," *IEEE TPAMI*, 2021.

[8] L. Ye, M. Rochan, Z. Liu, X. Zhang, and Y. Wang, "Referring segmentation in images and videos with cross-modal self-attention network," *IEEE TPAMI*, vol. 44, no. 7, pp. 3719–3732, 2021.

[9] T. Hui, S. Huang, S. Liu, Z. Ding, G. Li, W. Wang, J. Han, and F. Wang, "Collaborative spatial-temporal modeling for language-queried video actor segmentation," in *CVPR*, 2021, pp. 4187–4196.

[10] D. Li, R. Li, L. Wang, Y. Wang, J. Qi, L. Zhang, T. Liu, Q. Xu, and H. Lu, "You only infer once: Cross-modal meta-transfer for referring video object segmentation," in *AAAI*, 2022.

[11] W. Zhao, K. Wang, X. Chu, F. Xue, X. Wang, and Y. You, "Modeling motion with multi-modal features for text-based video segmentation," in *CVPR*, 2022, pp. 11 737–11 746.

[12] Z. Ding, T. Hui, J. Huang, X. Wei, J. Han, and S. Liu, "Language-bridged spatial-temporal interaction for referring video object segmentation," in *CVPR*, 2022, pp. 4964–4973.

[13] D. Wu, X. Dong, L. Shao, and J. Shen, "Multi-level representation learning with semantic alignment for referring video object segmentation," in *CVPR*, 2022, pp. 4996–5005.

[14] X. Yang, H. Wang, D. Xie, C. Deng, and D. Tao, "Object-agnostic transformers for video referring segmentation," *IEEE TIP*, vol. 31, pp. 2839–2849, 2022.

[15] H. Ding, C. Liu, S. Wang, and X. Jiang, "Vlt: Vision-language transformer and query generation for referring segmentation," *IEEE TPAMI*, 2022.

[16] C. Liang, W. Wang, T. Zhou, J. Miao, Y. Luo, and Y. Yang, "Local-global context aware transformer for language-guided video," *IEEE TPAMI*, 2023.

[17] A. Botach, E. Zheltonozhskii, and C. Baskin, "End-to-end referring video object segmentation with multimodal transformers," in *CVPR*, 2022.

[18] J. Wu, Y. Jiang, P. Sun, Z. Yuan, and P. Luo, "Language as queries for referring video object segmentation," in *CVPR*, 2022.

[19] X. Li, J. Wang, X. Xu, X. Li, Y. Lu, and B. Raj, "Rˆ 2vos: Robust referring video object segmentation via relational multimodal cycle consistency," *arXiv preprint arXiv:2207.01203*, 2022.

[20] A. Khoreva, A. Rohrbach, and B. Schiele, "Video object segmentation with language referring expressions," in *ACCV*, 2018, pp. 123–141.

[21] S. Seo, J.-Y. Lee, and B. Han, "Urvos: Unified referring video object segmentation network with a large-scale benchmark," in *ECCV*, 2020, pp. 208–223.

[22] F. Chen, D. Zhang, M. Han, X. Chen, J. Shi, S. Xu, and B. Xu, "Vlp: A survey on vision-language pre-training," *arXiv preprint arXiv:2202.09061*, 2022.

[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[24] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang *et al.*, "Grounded language-image pre-training," in *CVPR*, 2022, pp. 10 965–10 975.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[27] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020, pp. 213–229.

[28] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: deformable transformers for end-to-end object detection," in *ICLR*, 2021.

[29] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," in *CVPR*, 2021, pp. 8741–8750.

[30] M. Tang, Z. Wang, Z. Liu, F. Rao, D. Li, and X. Li, "Clip4caption: Clip for video caption," in *ACM MM*, 2021, pp. 4858–4862.

[31] H. Song, L. Dong, W. Zhang, T. Liu, and F. Wei, "Clip models are few-shot learners: Empirical studies on vqa and visual entailment," in *ACL*, 2022, pp. 6088–6100.

[32] Y. Zeng, X. Zhang, and H. Li, "Multi-grained vision language pre-training: Aligning texts with visual concepts," in *ICML*, 2022.

[33] L. Yao, R. Huang, L. Hou, G. Lu, M. Niu, H. Xu, X. Liang, Z. Li, X. Jiang, and C. Xu, "Filip: Fine-grained interactive language-image pre-training," in *ICLR*, 2022.

[34] H. Zhang, P. Zhang, X. Hu, Y.-C. Chen, L. H. Li, X. Dai, L. Wang, L. Yuan, J.-N. Hwang, and J. Gao, "Glipv2: Unifying localization and vision-language understanding," in *NeurIPS*, 2022.

[35] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 9992–10 002.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019.

[37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, vol. 32, 2019.

[38] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *ECCV*, 2016, pp. 69–85.

[39] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *CVPR*, 2016, pp. 11–20.

[40] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

Mingqi Gao is a Ph.D. candidate with the University of Warwick and Southern University of Science and Technology (SUSTech). His research interests include computer vision and video analysis.

Jinyu Yang is a Ph.D. candidate with the University of Birmingham and Southern University of Science and Technology (SUSTech). Her research interests include computer vision and object tracking.

Jungong Han is Chair Professor in Computer Vision at the Department of Computer Science, University of Sheffield, U.K. He also holds an Honorary Professorship with the University of Warwick, U.K. His research interests include computer vision, artificial intelligence, and machine learning.

Ke Lu is Professor with the University of Chinese Academy of Sciences. He is also with the Peng Cheng Laboratory. His current research areas include computer vision, 3D image reconstruction, and computer graphics.

Feng Zheng is Associate Professor with the Department of Computer Science and Engineering, Southern University of Science and Technology (SUSTech), China. His research interests include machine learning, computer vision, and human-computer interaction.

Giovanni Montana is Professor of Data Science at the University of Warwick where he holds a joint appointment with the Department of Statistics and Warwick Manufacturing Group (WMG). His research interests include data science, machine learning, and digital healthcare.