Context-Guided Spatio-Temporal Video Grounding

University of Chinese Academy of Sciences

ISCAS UNT

Xin Gu[1,3*]   **Heng Fan**[2*]   **Yan Huang**[2]   **Tiejian Luo**[1]   **Libo Zhang**[1,3†] (*equal contributions, †Corresponding author)

[1]University of Chinese Academy of Sciences   [2]University of North Texas   [3]Institute of Software Chinese Academy of Sciences

CVPR SEATTLE, WA JUNE 17-21, 2024

Spatio-Temporal Video Grounding

- **What is spatio-temporal video grounding (STVG)?**
  – STVG aims to localize the object of interest in an untrimmed video with a spatio-temporal tube given a free-form textual query

Input text query:
What does the adult ride in the playground?

Output:
A spatio-temporal tube

t_{start} t_{end} (Yang et al, CVPR'2022)
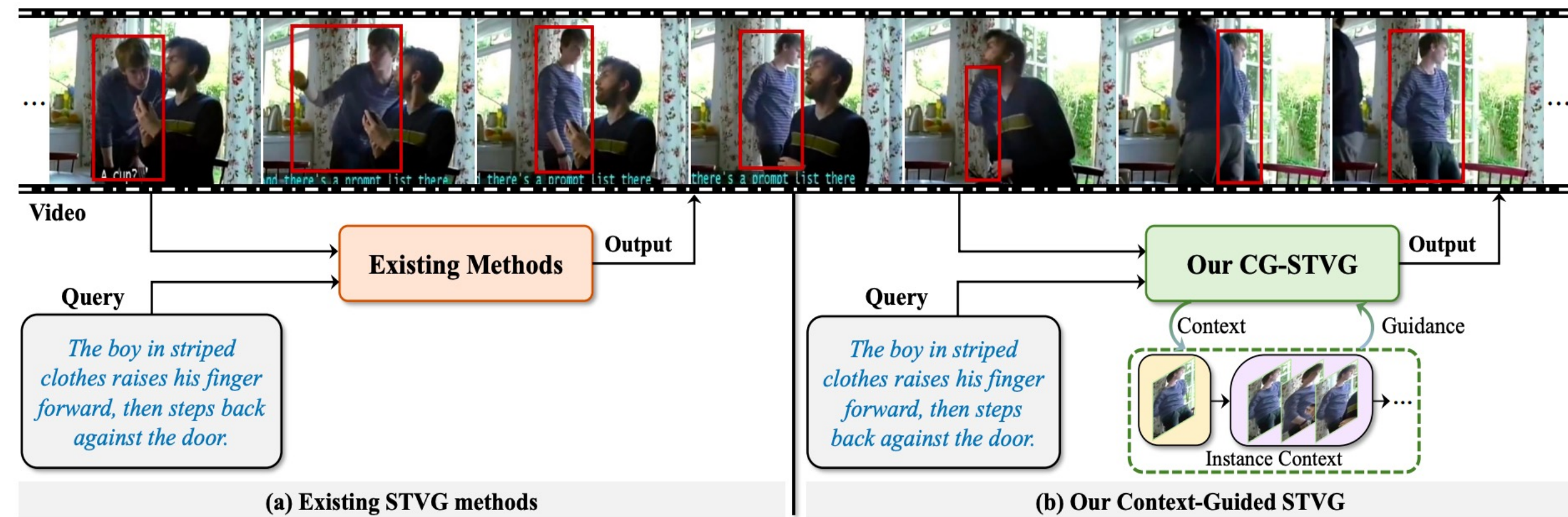
spatio-temporal tube
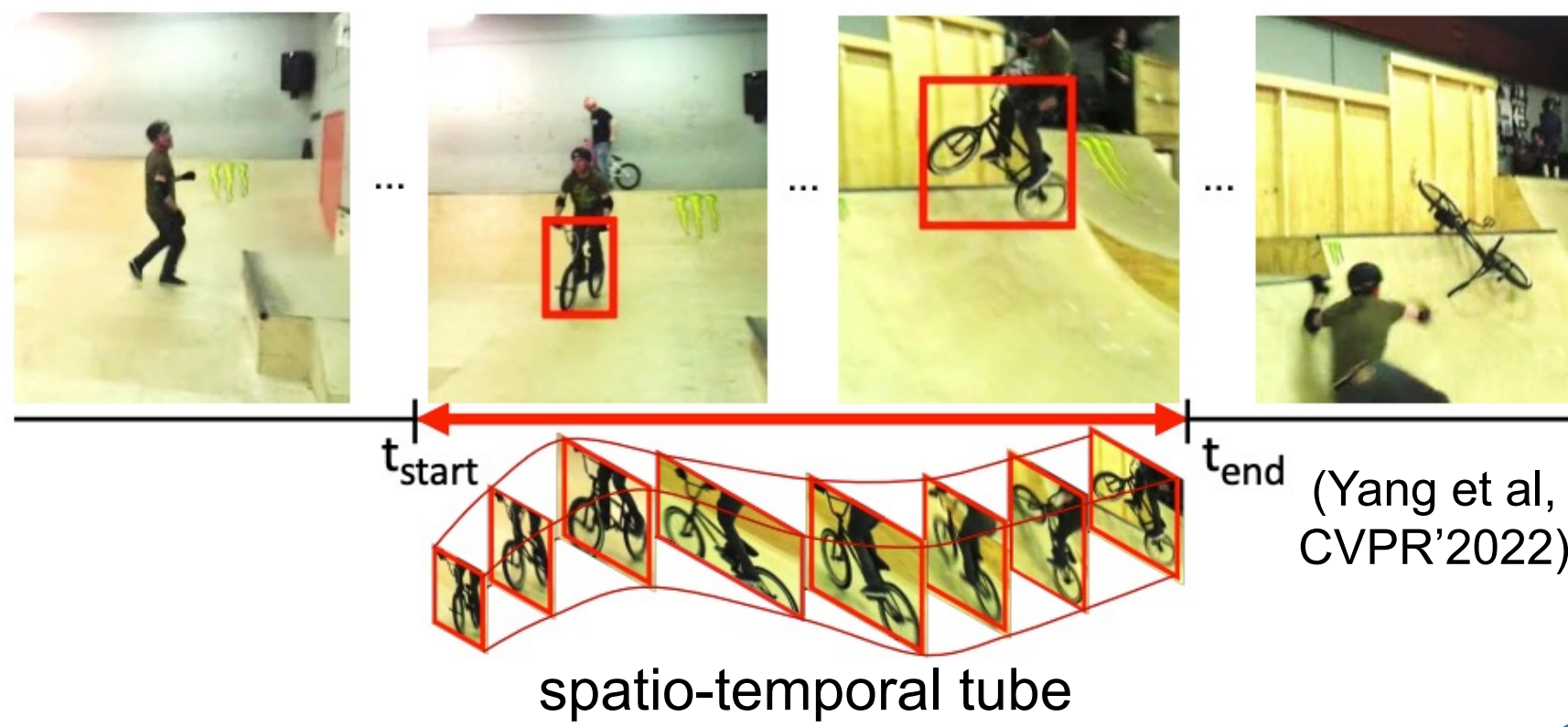
Motivation

Figure 1: Comparison of existing methods (a) with our context-guided STVG (b)

- **Existing STVG Methods (Fig. 1 (a))**
  - Text query as the *only* cue for target localization
  - *Insufficient* to distinguish foreground object in complex scenes
  - Enhance text with more fine-grained information: (i) *laborious*; (ii) *more computational overheads*; (iii) *still difficult to describe visual details*.

- **Our context-guided STVG (Fig. 1 (b))**
  - A famous adage: "*A Picture Is Worth a Thousand Words*"
  - Exploit visual information of the object to offer a guidance, directly from the vision perspective, for improving STVG
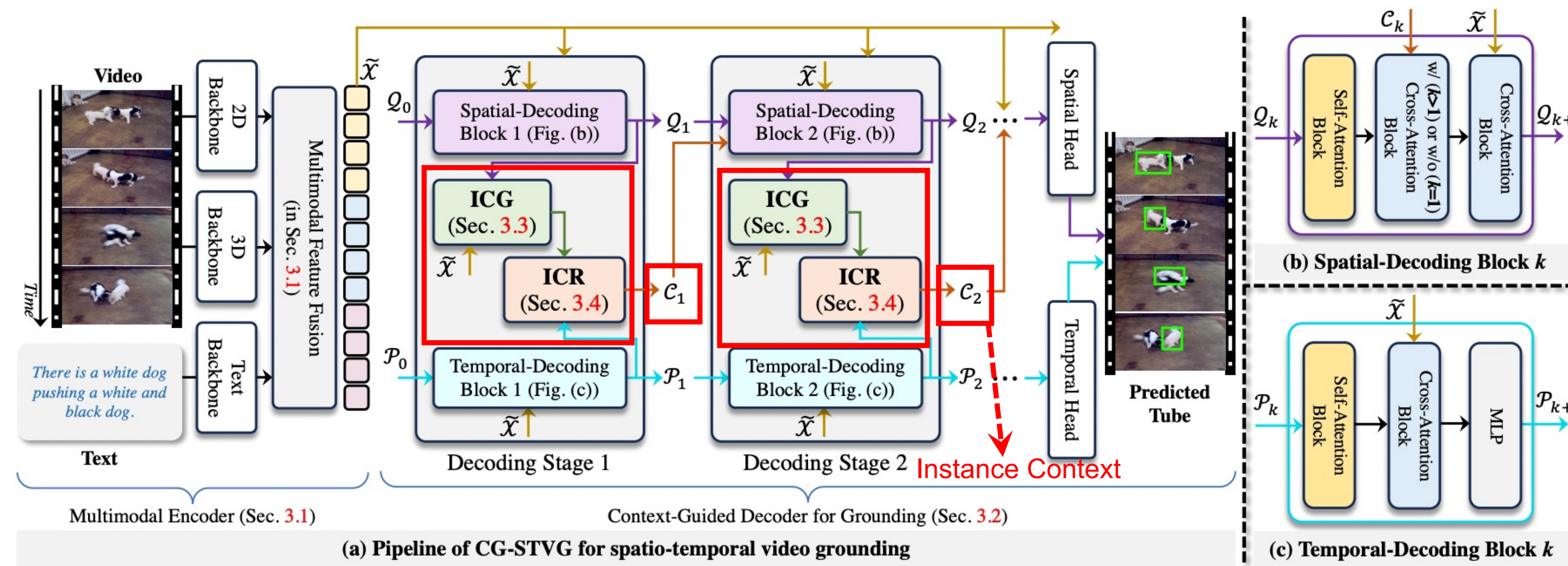
The Proposed Methodology

Figure 2: Overview of the proposed context-guided spatio-temporal video grounding

- **Context-guided spatio-temporal video grounding:** Mining instance visual context from the video to guide spatio-temporal target localization (Fig. 2)
  - Feature extraction and interaction for video (2D appearance and 3D motion features) and text
  - Spatial- and temporal-decoding for target localization
  - Instance visual context is mined during decoding, via ICG and ICR, and used for guiding localization

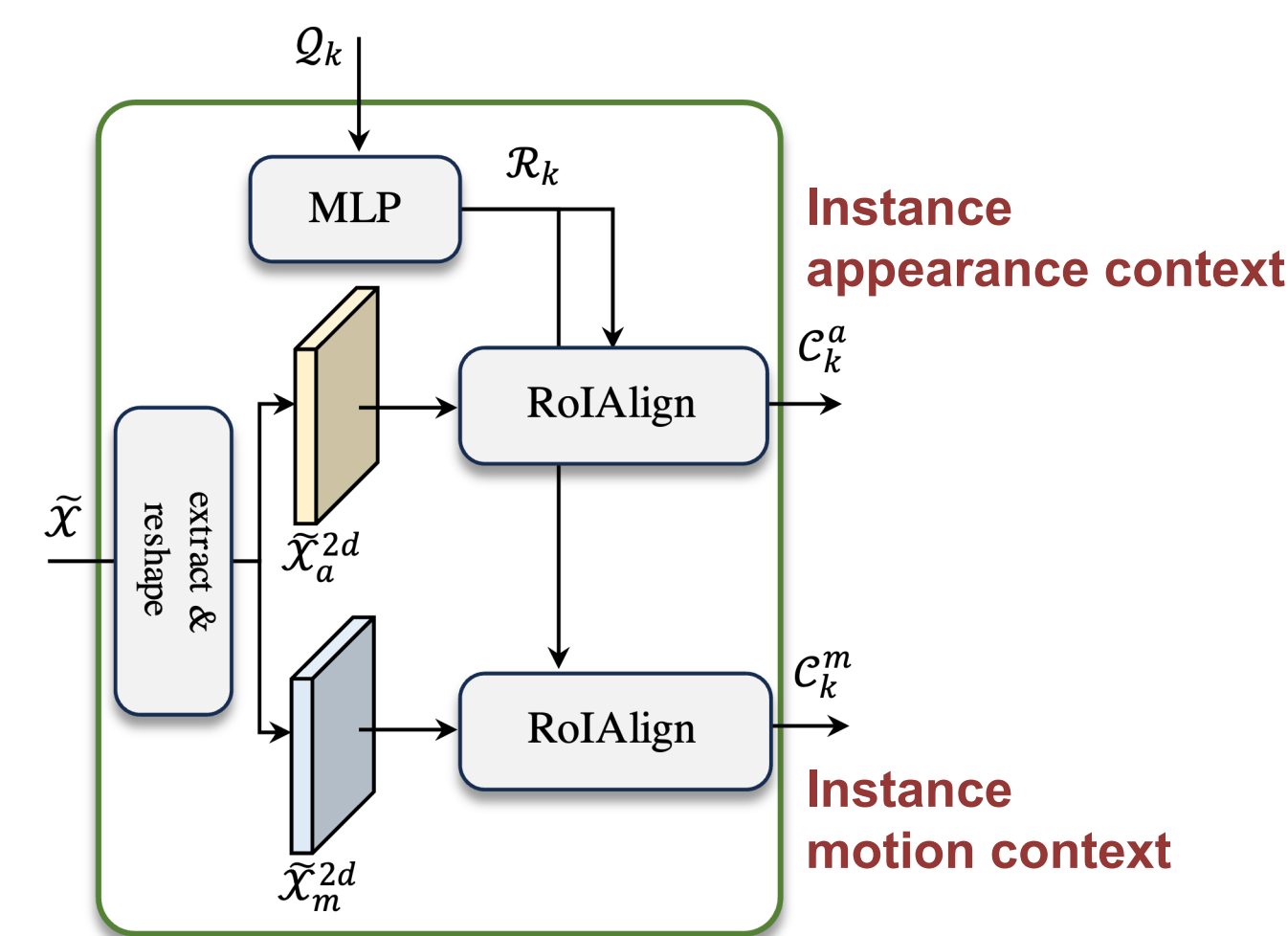- **Core modules:** ICG for instance context generation and ICR instance context refinement

Figure 3: Illustration of ICG

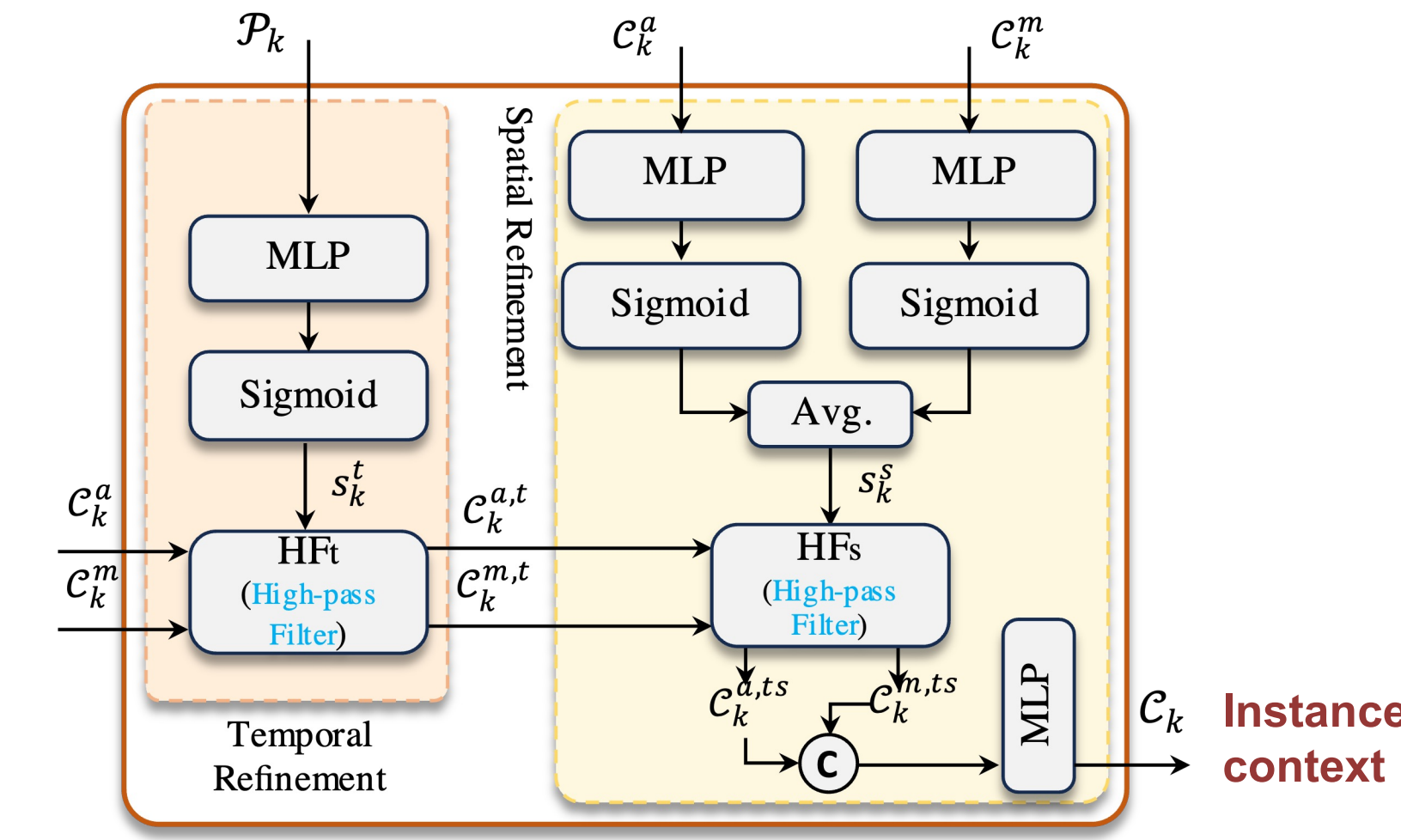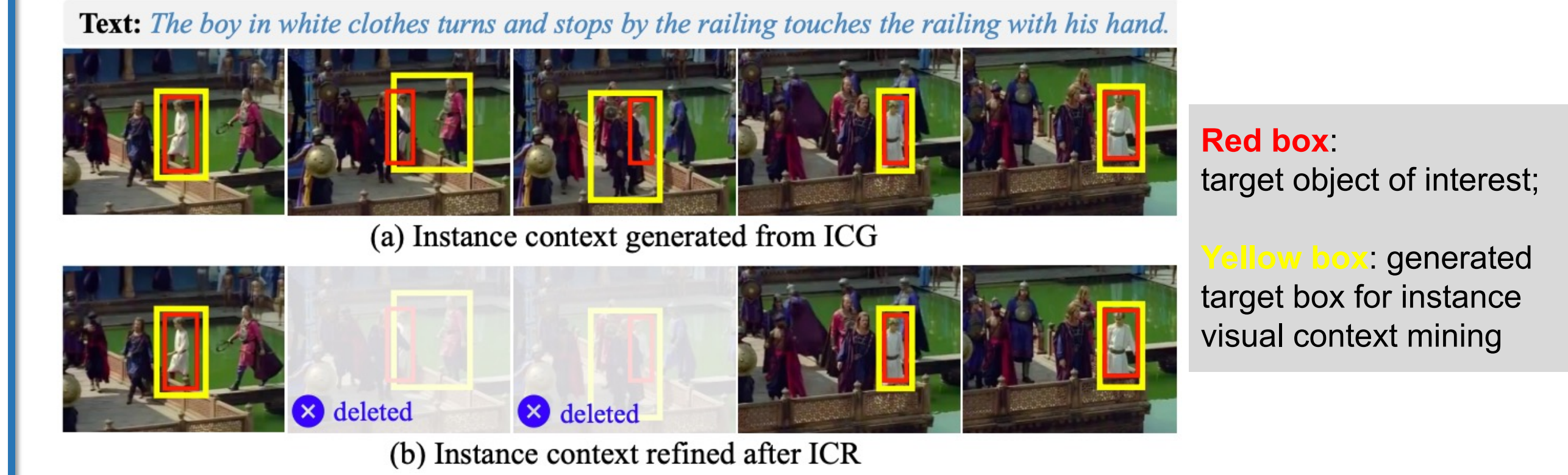**Figure 4**: Illustration of ICR

Analysis

- **Illustration of ICG and ICR**

Text: *The boy in white clothes turns and stops by the railing touches the railing with his hand.*

(a) Instance context generated from ICG

(b) Instance context refined after ICR

Red box: target object of interest;

Yellow box: generated target box for instance visual context mining

- **Attention maps**

without instance context

with instance context

Text: *The bald man waves his hat, then turns and walks away.*

Red box: target object of interest;

Experiments

Methods	m.tIoU	m.vIoU	vIoU@0.3	vIoU@0.5
STGVT [TCSV721] [32]	-	18.2	26.6	9.5
STVGBert [ICCV2021] [30]	-	20.4	29.4	11.3
TubeDETR [CVPR22] [36]	43.7	32.4	49.8	23.5
STCAT [NeurIPS22] [17]	49.4	35.1	57.7	30.1
CSDVL [CVPR23] [22]	-	36.9	62.2	34.8
Baseline	50.4	36.5	58.6	32.3
CG-STVG	52.8 (+2.4)	38.4 (+1.9)	61.5 (+2.9)	36.3 (+4.0)

Table: Results on HCSTVG-v1 test set (%)

Methods	m.tIoU	m.vIoU	vIoU@0.3	vIoU@0.5
PCC [arxiv21] [8]	-	30.0	-	-
2D-Tan [arxiv2021] [31]	-	30.4	50.4	18.8
MMN [AAAI22] [35]	-	30.3	49.0	25.6
TubeDETR [CVPR22] [36]	-	36.4	58.8	30.6
CSDVL [CVPR23] [22]	58.1	38.7	65.5	33.8
Baseline	58.6	37.8	62.4	32.1
CG-STVG	60.0 (+1.4)	39.5 (+1.7)	64.5 (+2.1)	36.3 (+4.2)

Table: Results on HCSTVG-v2 test set (%)

Methods	Declarative Sentences				Interrogative Sentences			
	m.tIoU	m.vIoU	vIoU@0.3	vIoU@0.5	m.tIoU	m.vIoU	vIoU@0.3	vIoU@0.5
STGRN [CVPR20] [43]	48.5	19.8	25.8	14.6	47.0	18.3	21.1	12.8
OMRN [IJCAI20] [41]	50.7	23.1	32.6	16.4	49.2	20.6	28.4	14.1
STGVT [TCSV721] [32]	-	21.6	29.8	18.9	-	-	-	-
STVGBert [ICCV21] [30]	-	24.0	30.9	18.4	-	22.5	26.0	16.0
TubeDETR [CVPR22] [36]	48.1	30.4	42.5	28.2	46.9	25.7	35.7	23.2
STCAT [NeurIPS23] [17]	50.8	33.1	46.2	32.6	49.7	28.2	39.2	26.6
CSDVL [CVPR23] [22]	-	33.7	47.2	32.8	-	28.5	39.9	26.2
Baseline	49.7	32.4	45.0	31.4	48.8	27.7	38.7	25.6
CG-STVG	51.4 (+1.7)	34.0 (+1.6)	47.7 (+2.7)	33.1 (+1.7)	49.9 (+1.1)	29.0 (+1.3)	40.5 (+1.8)	27.5 (+1.9)

Visual context has significantly improved the performance!

Table: Results on VidSTG test set (%)

Code/Results