Scan2CapMMT

Dense Captioning for 3D Scenes with Transformers

Scan2CapMMT

I. Scan2CapMMT Recap

II. Improving Scan2CapMMT

III. Quantitative & Qualitative Results

IV. Detection with Transformers

V. Timeline until the Final Presentation

Scan2CapMMT

I. Scan2CapMMT Recap

II. Improving Scan2CapMMT

III. Quantitative & Qualitative Results

IV. Detection with Transformers

V. Timeline until the Final Presentation

I. Scan2Cap Recap

Point Cloud

Object Detection Module

Relational Graph Module

Captioning Module

Captions for the Object Proposals

I. Scan2Cap Recap

Point Cloud

PointNet++

Voting Module

Proposal Module

Object Proposals with Features

Object Masks

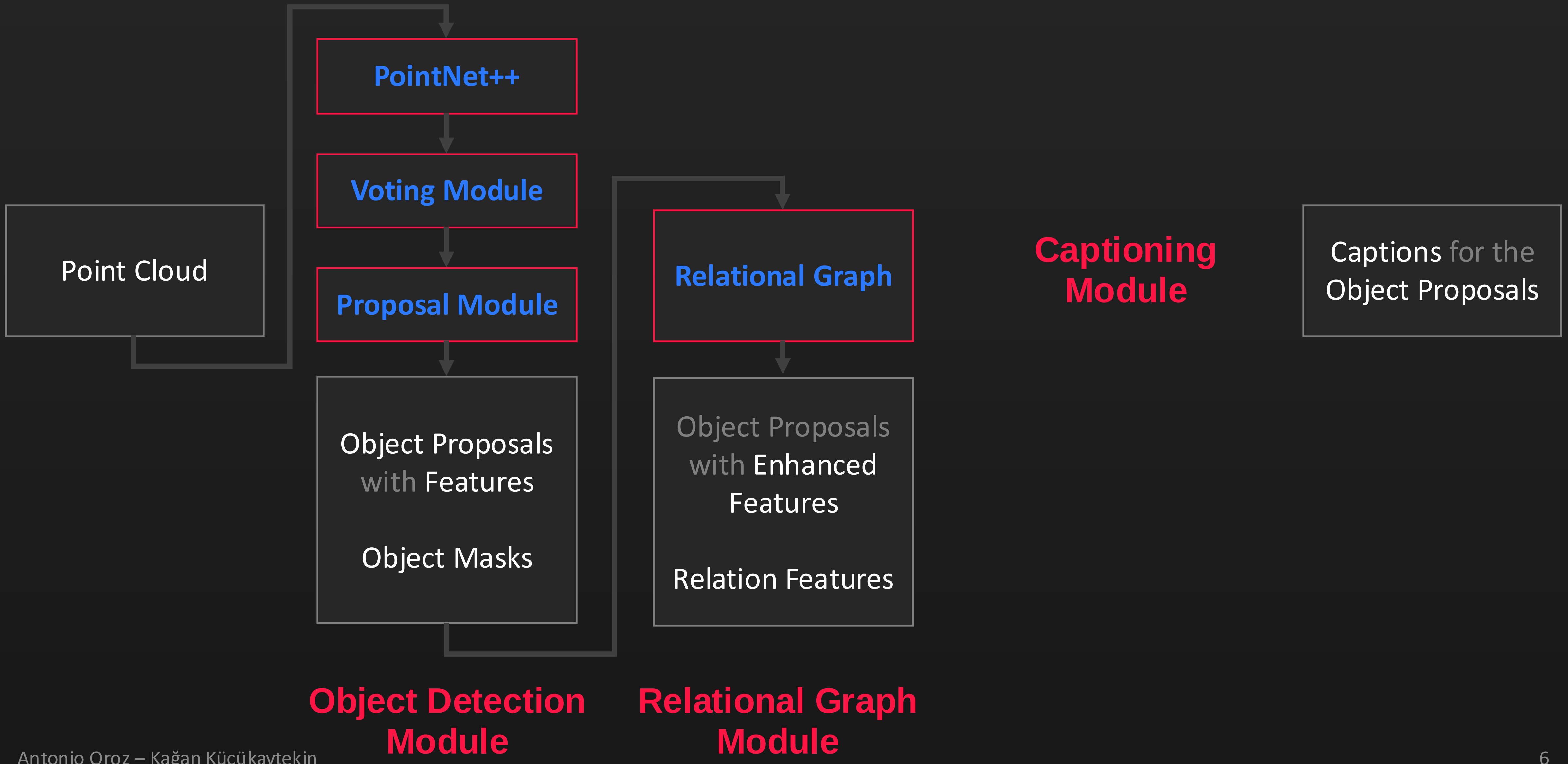
Object Detection Module

Relational Graph Module

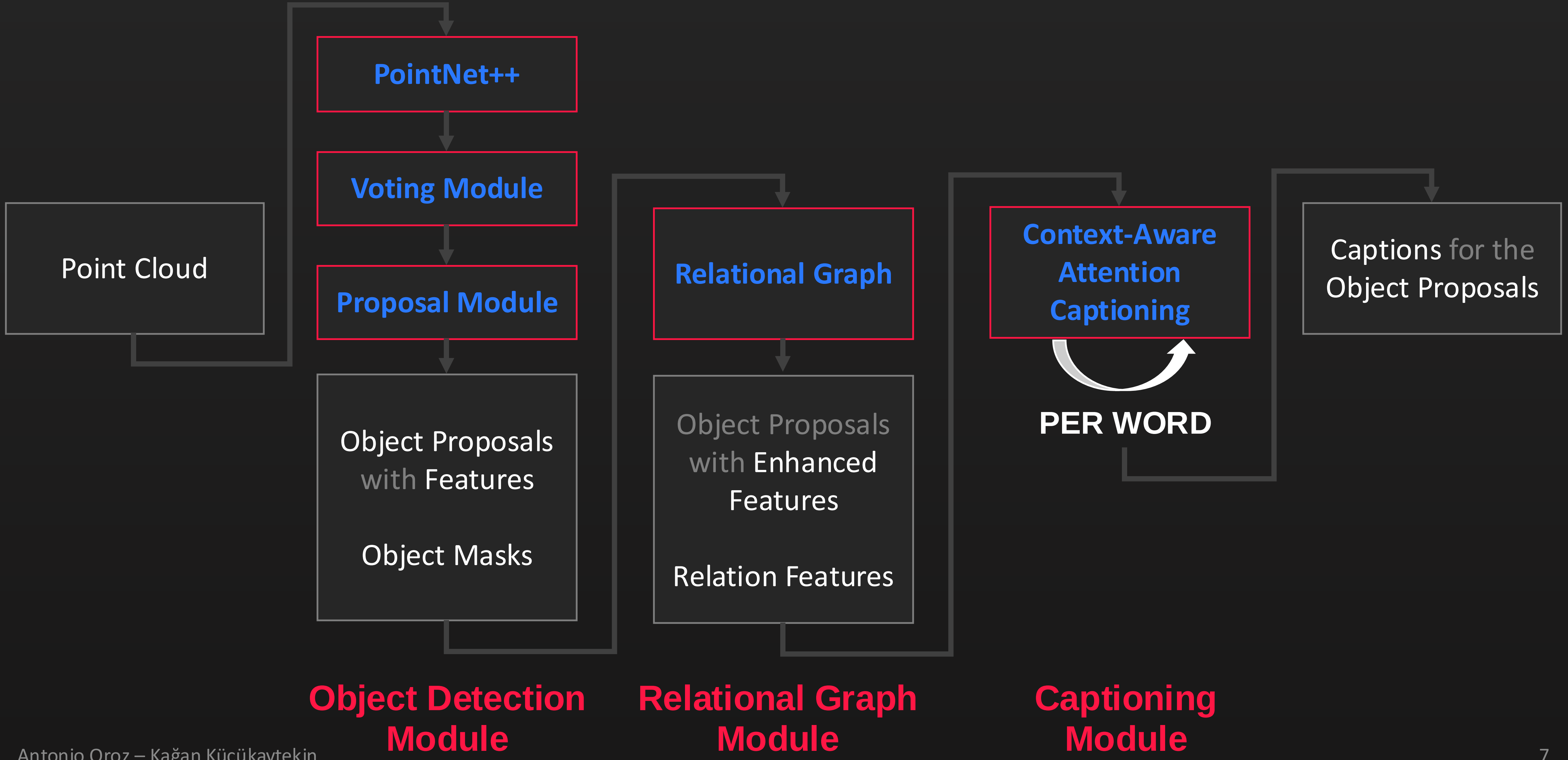
Captioning Module

Captions for the Object Proposals

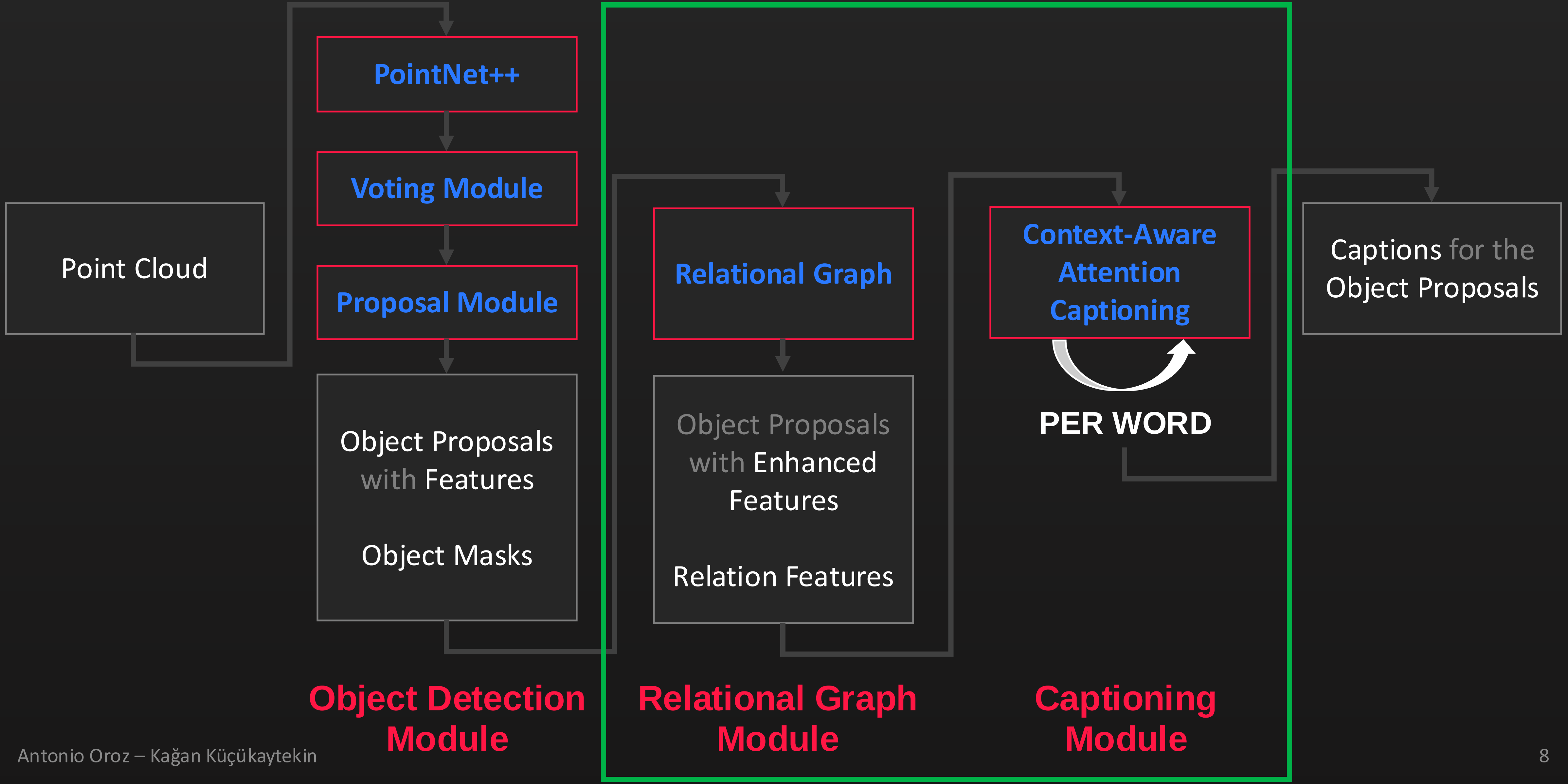
I. Scan2Cap Recap



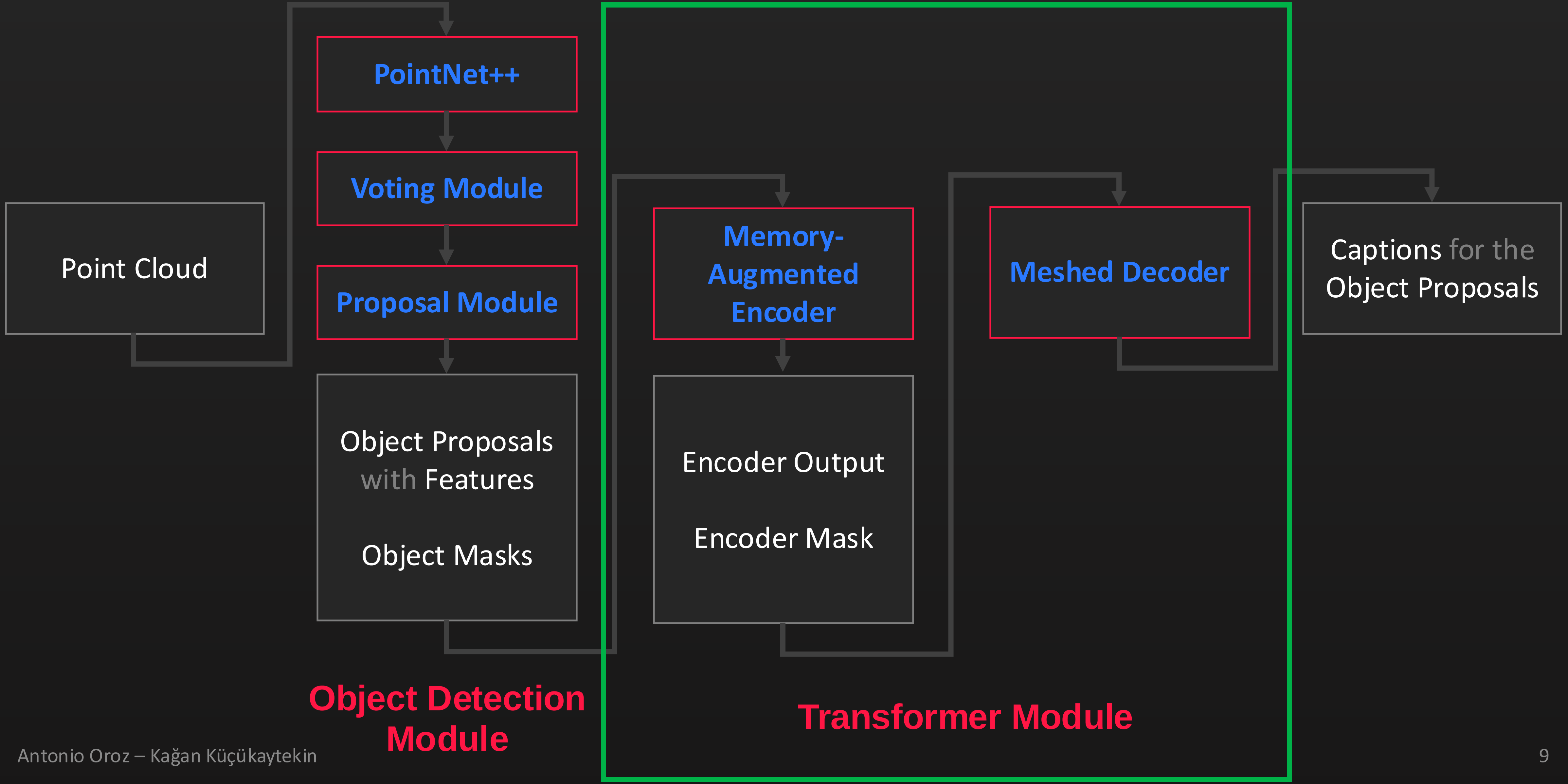
I. Scan2Cap Recap



I. Scan2Cap Recap



I. Scan2CapMMT Recap



Scan2CapMMT

I. Scan2CapMMT Recap

II. Improving Scan2CapMMT

III. Quantitative & Qualitative Results

IV. Detection with Transformers

V. Timeline until the Final Presentation

Scan2CapMMT

I. Scan2CapMMT Recap

II. Improving Scan2CapMMT

III. Quantitative & Qualitative Results

IV. Detection with Transformers

V. Timeline until the Final Presentation

II. Improving Scan2CapMMT

Beam Search

ITERATIVE SEARCH

A	CHAIR

II. Improving Scan2CapMMT

Beam Search

ITERATIVE SEARCH

A | CHAIR | NEXT: 0.5

INBETWEEN: 0.3

BEHIND: 0.1

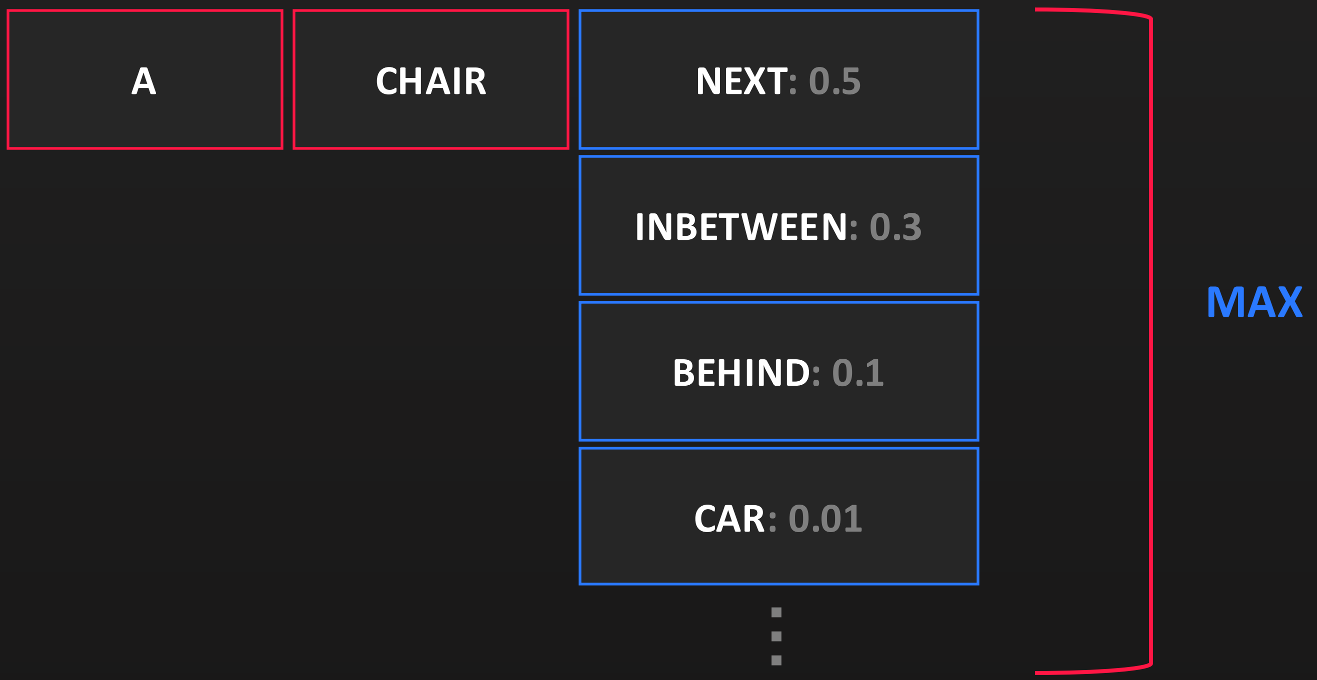
CAR: 0.01

⋮

II. Improving Scan2CapMMT

Beam Search

ITERATIVE SEARCH



II. Improving Scan2CapMMT

Beam Search

ITERATIVE SEARCH

A	CHAIR	NEXT: 0.5	...

II. Improving Scan2CapMMT

Beam Search

A: 0.9	CHAIR: 0.4

BEAM SEARCH
SIZE 2

A: 0.9	TABLE: 0.3

II. Improving Scan2CapMMT

Beam Search

BEAM SEARCH
SIZE 2

A: 0.9 | CHAIR: 0.4 | NEXT: 0.5
INBETWEEN: 0.3
BEHIND: 0.1
⋮

A: 0.9 | TABLE: 0.3 | NEXT: 0.3
INBETWEEN: 0.2
⋮

II. Improving Scan2CapMMT

Beam Search

BEAM SEARCH
SIZE 2

A: 0.9	CHAIR: 0.4	NEXT: 0.5	0.9 * 0.4 * 0.5 = 0.18
		INBETWEEN: 0.3	0.9 * 0.4 * 0.3 = 0.108
		BEHIND: 0.1	0.9 * 0.4 * 0.1 = 0.036
		⋮	
A: 0.9	TABLE: 0.3	NEXT: 0.3	0.9 * 0.3 * 0.3 = 0.081
		INBETWEEN: 0.2	0.9 * 0.3* 0.2 = 0.054
		⋮	

II. Improving Scan2CapMMT

Beam Search

BEAM SEARCH
SIZE 2

A: 0.9	CHAIR: 0.4	NEXT: 0.5	0.9 * 0.4 * 0.5 = 0.18
		INBETWEEN: 0.3	0.9 * 0.4 * 0.3 = 0.108
		BEHIND: 0.1	0.9 * 0.4 * 0.1 = 0.036
		⋮	
A: 0.9	TABLE: 0.3	NEXT: 0.3	0.9 * 0.3 * 0.3 = 0.081
		INBETWEEN: 0.2	0.9 * 0.3* 0.2 = 0.054
		⋮	

MAX

II. Improving Scan2CapMMT

Beam Search

BEAM SEARCH SIZE 2

A: 0.9	CHAIR: 0.4	NEXT: 0.5	0.9 * 0.4 * 0.5 = 0.18
		INBETWEEN: 0.3	0.9 * 0.4 * 0.3 = 0.108
		BEHIND: 0.1	0.9 * 0.4 * 0.1 = 0.036
		⋮	
A: 0.9	TABLE: 0.3	NEXT: 0.3	0.9 * 0.3 * 0.3 = 0.081
		INBETWEEN: 0.2	0.9 * 0.3* 0.2 = 0.054
		⋮	

MAX

II. Improving Scan2CapMMT

Beam Search

A: 0.9	CHAIR: 0.4	NEXT: 0.5	...

BEAM SEARCH
SIZE 2

A: 0.9	CHAIR: 0.4	INBETWEEN: 0.3	...

II. Improving Scan2CapMMT

Beam Search | Reinforcement Learning

	CIDEr
This is a white sink…	0.41
This white a rectangular…	0.32
This is kitchen white…	0.24
This white a to sink…	0.001
This is is white oven…	0.09

II. Improving Scan2CapMMT

Beam Search

Reinforcement Learning

	CIDEr
This is a white sink...	0.41
This white a rectangular...	0.32
This is kitchen white...	0.24
This white a to sink...	0.001
This is is white oven...	0.09

$$-\frac{1}{k}\sum_{i=1}^{k}\left(r\left(w^{i}\right)-b\right)\log\left(p\left(w^{i}\right)\right)$$

II. Improving Scan2CapMMT

Beam Search

Reinforcement Learning

	CIDEr
This is a white sink…	0.41
This white a rectangular…	0.32
This is kitchen white…	0.24
This white a to sink…	0.001
This is is white oven…	0.09

MEAN
b=0.21

$$-\frac{1}{k}\sum_{i=1}^{k}\left(r\left(w^{i}\right)-b\right)\log\left(p\left(w^{i}\right)\right)$$

II. Improving Scan2CapMMT

Beam Search

Reinforcement Learning

$r(w^i) - b$

This is a white sink…	0.2
This white a rectangular…	0.11
This is kitchen white…	0.03
This white a to sink…	-0.21
This is is white oven…	-0.12

MEAN
b=0.21

$$-\frac{1}{k}\sum_{i=1}^{k}\left(r\left(w^i\right) - b\right)\log\left(p\left(w^i\right)\right)$$

Scan2CapMMT

I. Scan2CapMMT Recap

II. Improving Scan2CapMMT

III. Quantitative & Qualitative Results

IV. Detection with Transformers

V. Timeline until the Final Presentation

Scan2CapMMT

III. Quantitative Results: vs Scan2Cap

@0.5IoU

Model	CIDEr	Bleu-4	Meteor	Rouge
VoteNet+GRU	**34.31**	21.42	20.13	41.33
VoteNet+MMT *=Scan2CapMMT*	32.99	**21.92**	**20.96**	**44.40**

III. Quantitative Results: vs Scan2Cap

@0.5IoU

Model	CIDEr	Bleu-4	Meteor	Rouge
VoteNet+CAC	**36.15**	21.58	20.65	41.78
VoteNet+MMT *=Scan2CapMMT*	32.99	**21.92**	**20.96**	**44.40**

III. Quantitative Results: vs Scan2Cap

@0.5IoU

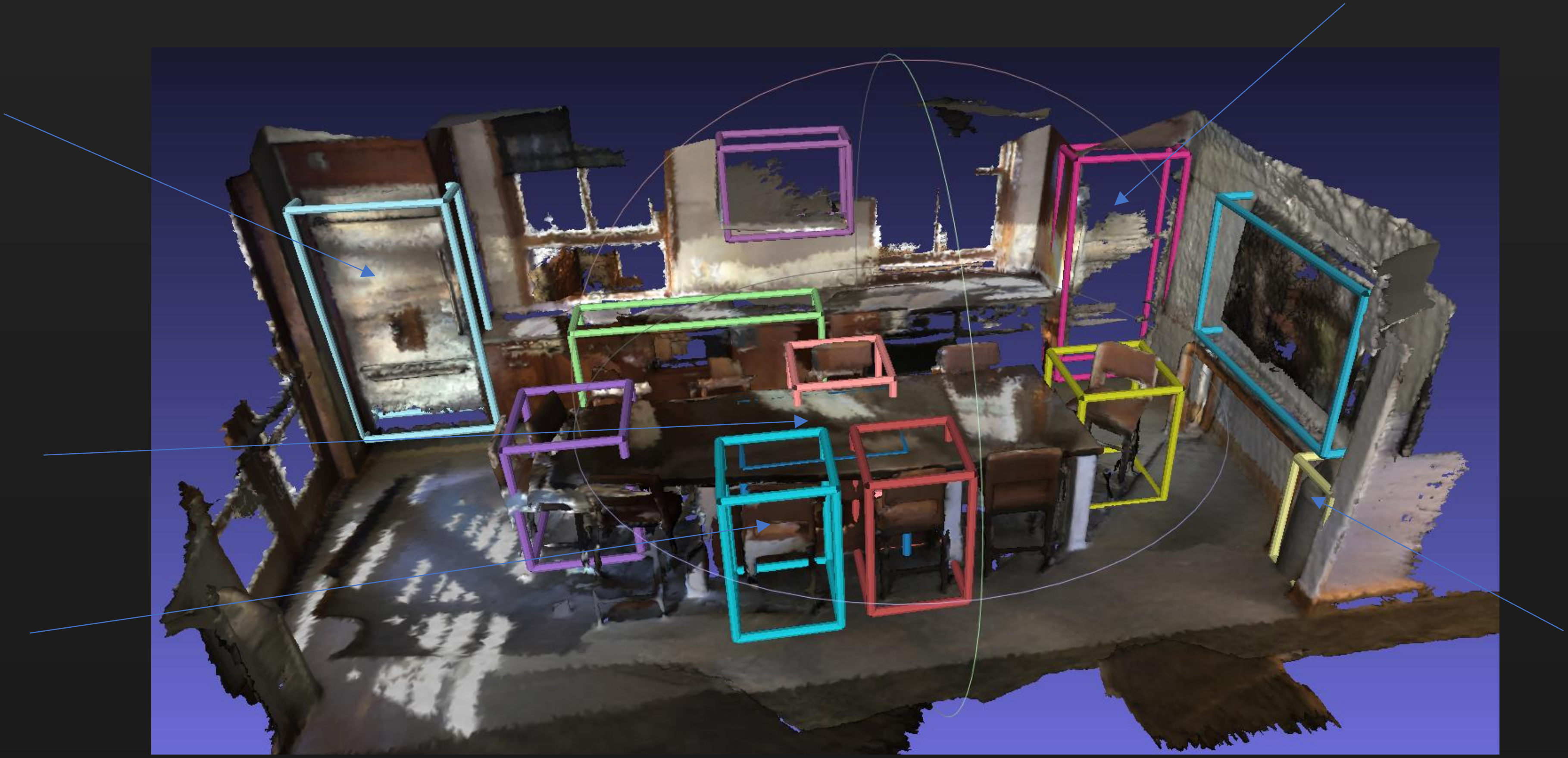
Model	CIDEr	Bleu-4	Meteor	Rouge
VoteNet+RG+CAC *= Scan2Cap*	**39.08**	**23.32**	**21.97**	**44.78**
Scan2CapMMT *= Scan2CapMMT*	32.99	21.92	20.96	44.40

III. Quantitative Results: Reinforcement Learning

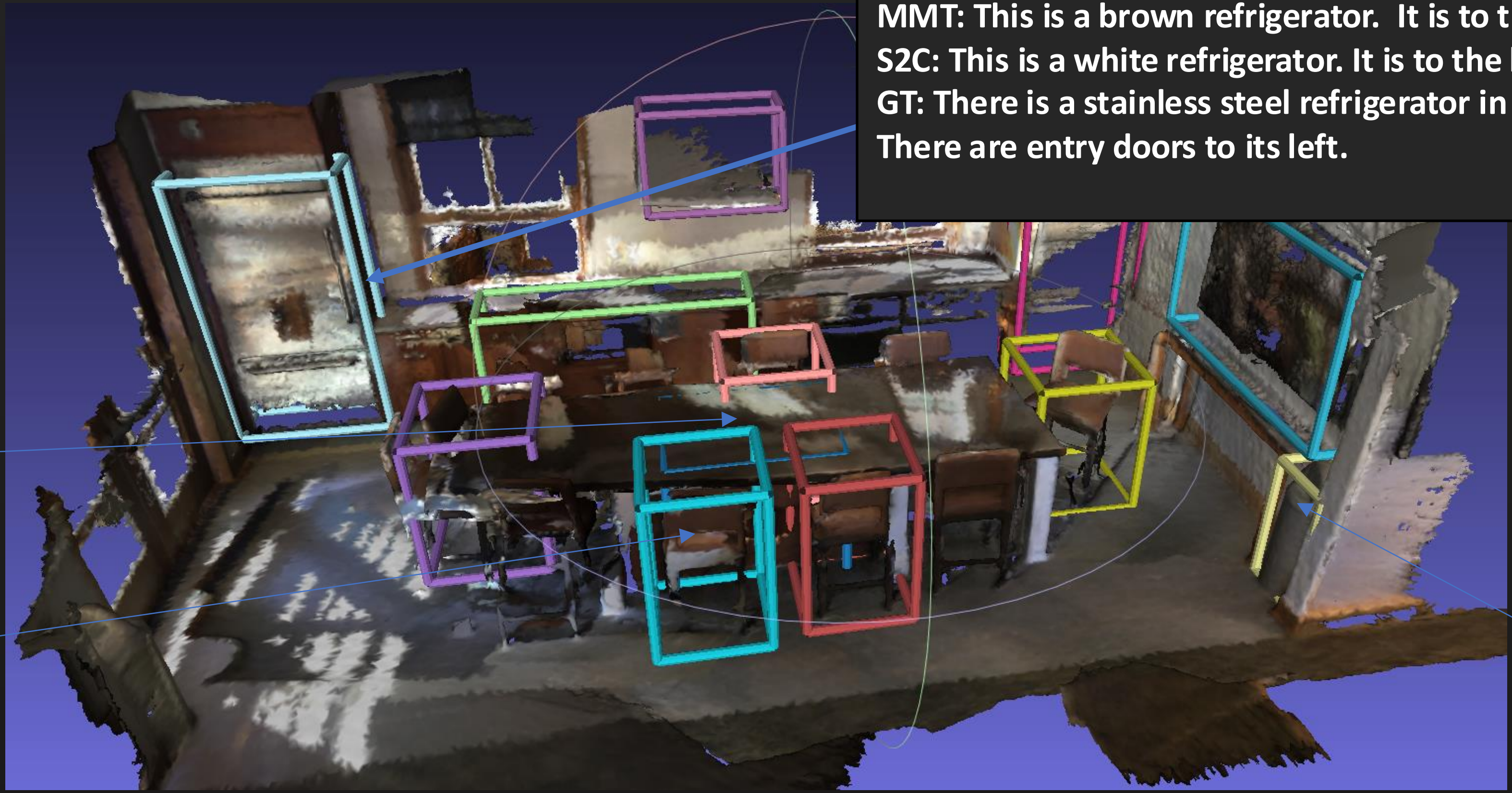
@0.5IoU

Model	CIDEr	Bleu-4	Meteor	Rouge
VoteNet+RG+CAC *= Scan2Cap*	**39.08**	23.32	**21.97**	**44.78**
VoteNet+MMT *= Scan2CapMMT*	32.99	21.92	20.96	44.40
Scan2CapMMT RL	36.18	**23.68**	21.33	44.64

III. Qualitative Results



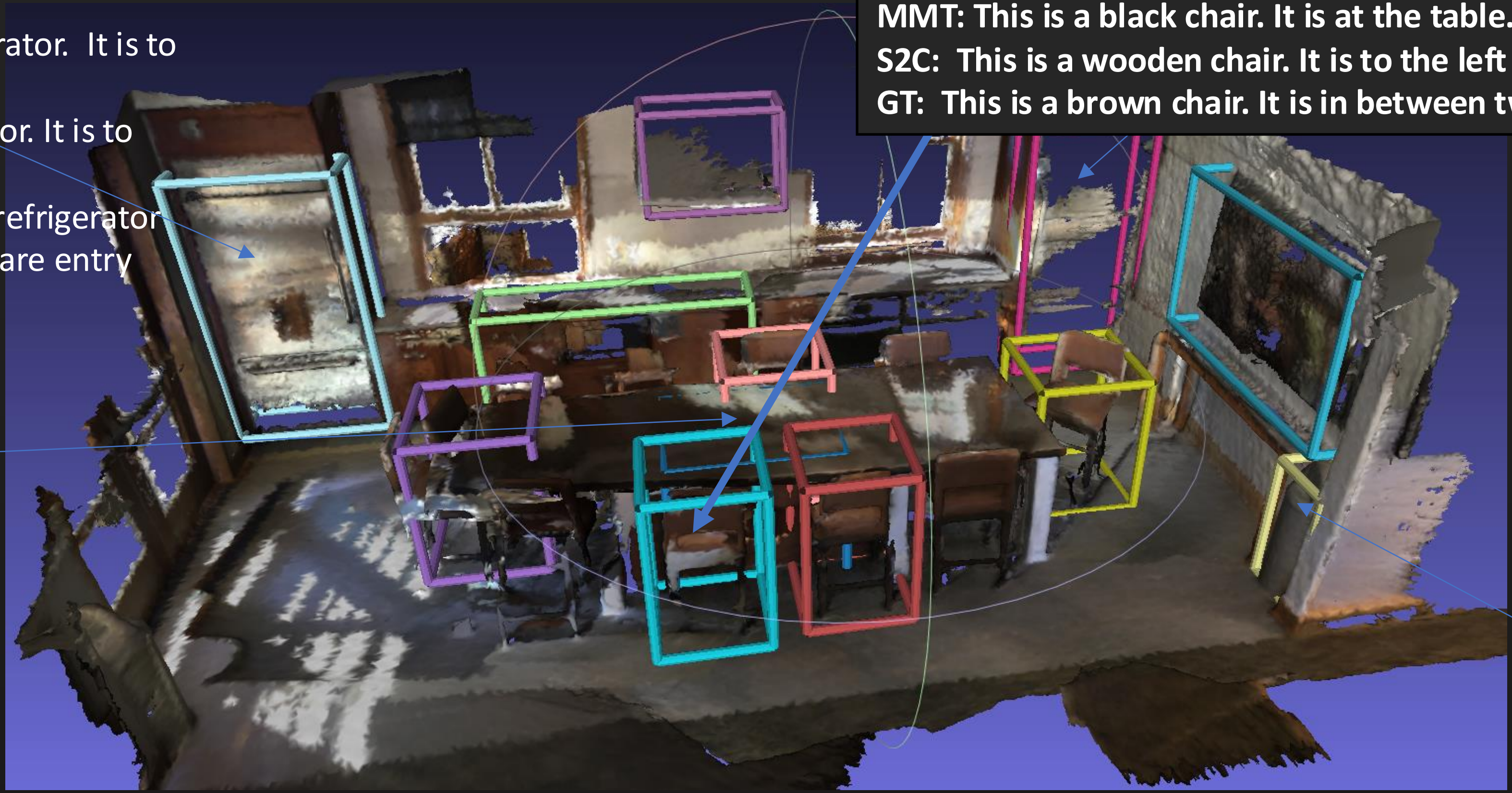
III. Qualitative Results



MMT: This is a brown refrigerator. It is to the left of the door.
S2C: This is a white refrigerator. It is to the left of the door.
GT: There is a stainless steel refrigerator in corner of the room. There are entry doors to its left.

III. Qualitative Results

MMT: This is a brown refrigerator. It is to the left of the door.
S2C: This is a white refrigerator. It is to the left of the door.
GT: There is a stainless steel refrigerator in corner of the room. There are entry doors to its left.

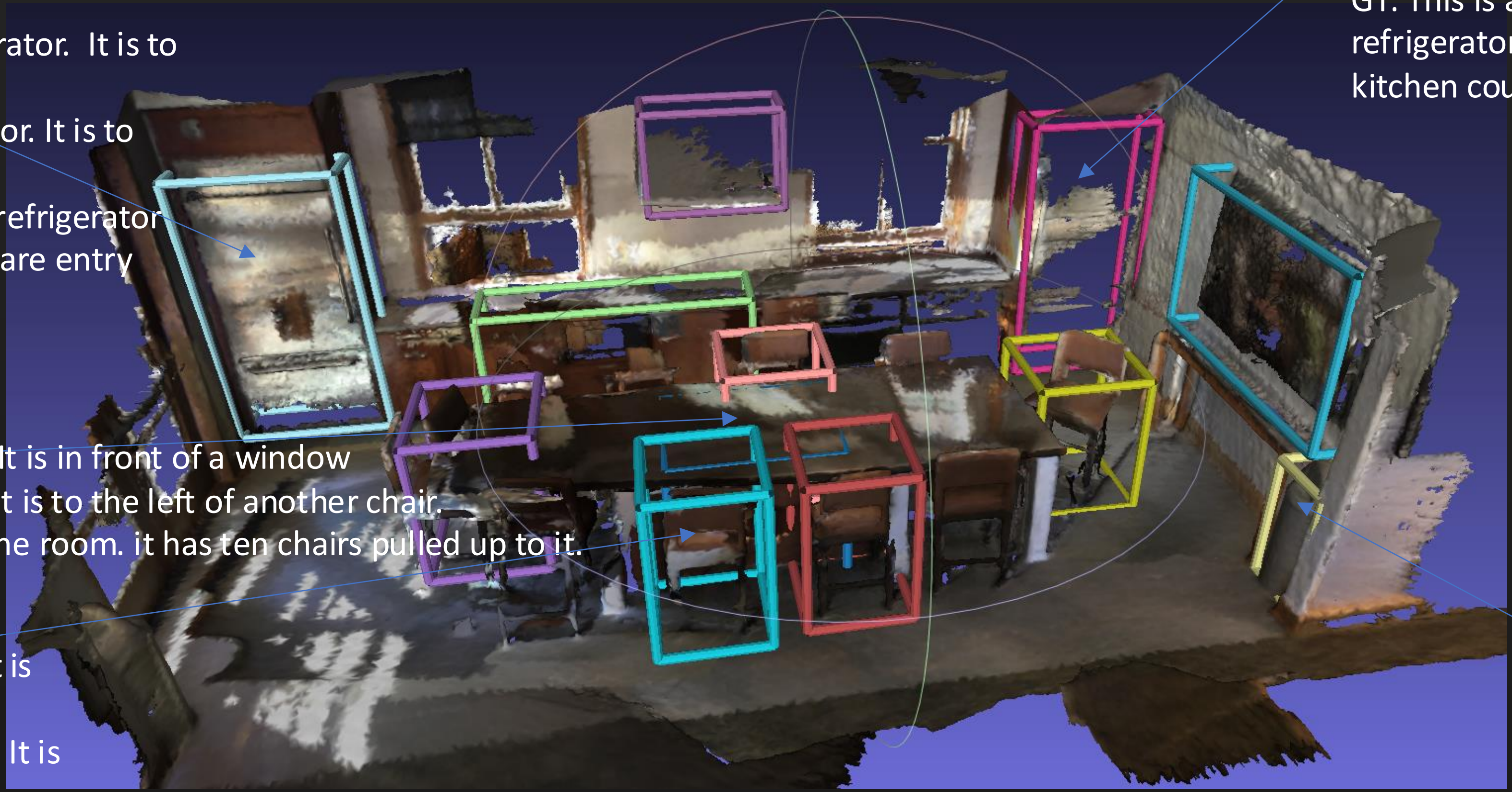


MMT: This is a black chair. It is at the table.
S2C: This is a wooden chair. It is to the left of another chair.
GT: This is a brown chair. It is in between two other chairs.

III. Qualitative Results

MMT: This is a brown door. It is to the right of the door.
S2C: This is a white door. It is to the left of the shelf.
GT: This is a stainless steel refrigerator. It is to the right of a kitchen counter.

MMT: This is a brown refrigerator. It is to the left of the door.
S2C: This is a white refrigerator. It is to the left of the door.
GT: There is a stainless steel refrigerator in corner of the room. There are entry doors to its left.

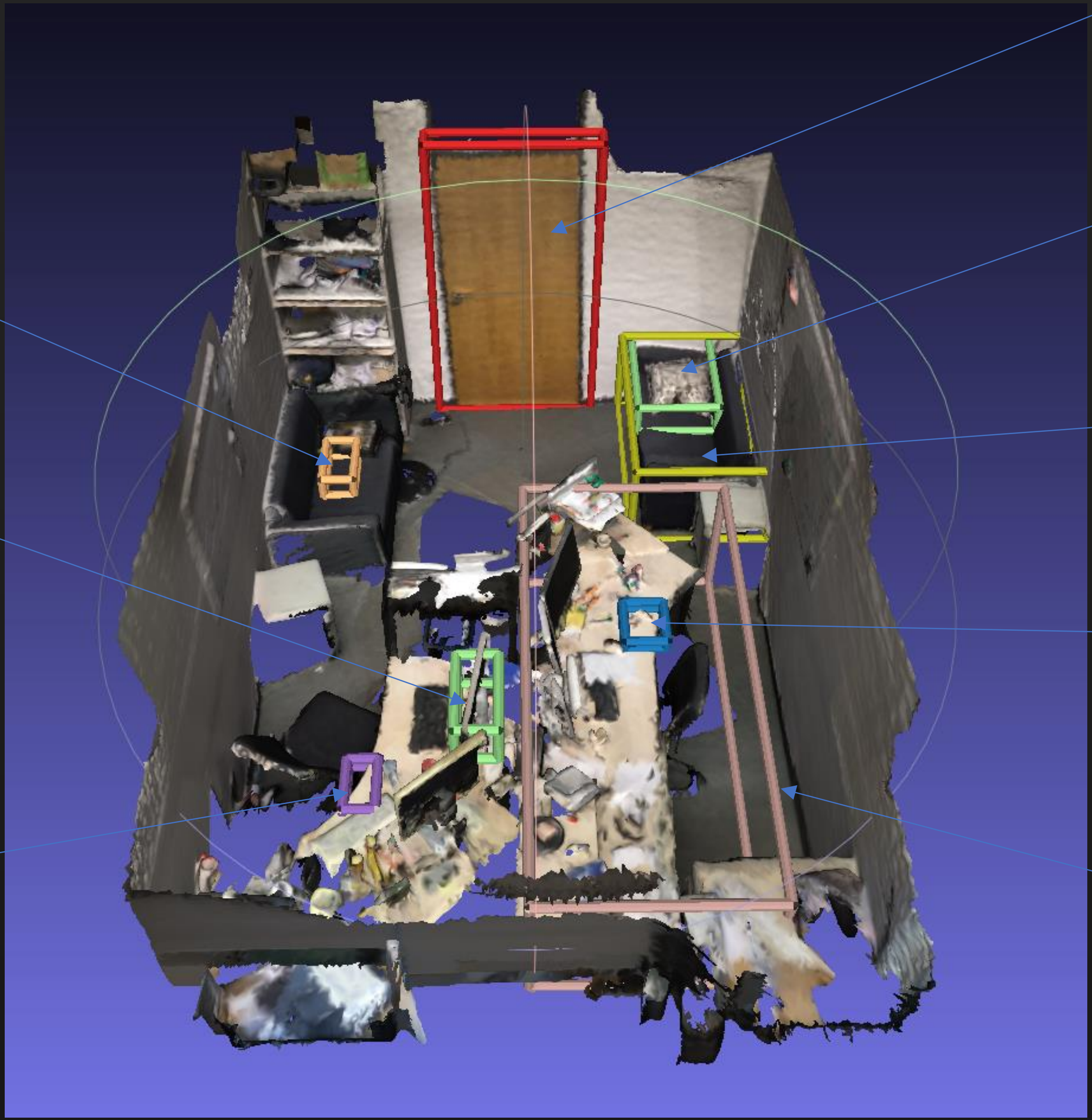


MMT: This is a brown table. It is in front of a window
S2C: This is a wooden chair. It is to the left of another chair.
GT: there is a large table in the room. it has ten chairs pulled up to it.

MMT: This is a black trash can. It is to the right of the door.
S2C: This is a trash can. It sets against the wall.
GT: This is a gray trash can. It is to the right of a table.

MMT: This is a black chair. It is at the table.
S2C: This is a wooden chair. It is to the left of another chair.
GT: This is a brown chair. It is in between two other chairs.

III. Qualitative Results



III. Qualitative Results

MMT: This is a brown door. It is to the right of a bookshelf.
S2C: This is a white door. It is to the left of a couch.
GT: A light brown door beside a tall shelf. A black couch is to the right of it .

III. Qualitative Results

MMT: This is a brown door. It is to the right of a bookshelf.
S2C: This is a white door. It is to the left of a couch.
GT: A light brown door beside a tall shelf. A black couch is to the right of it .

MMT: This is a black keyboard. It is on the desk.
S2C: This is a square pillow. It is on the couch.
GT: This is a small square gray pillow. It is located on a black couch.

III. Qualitative Results

MMT: This is a brown door. It is to the right of a bookshelf.
S2C: This is a white door. It is to the left of a couch.
GT: A light brown door beside a tall shelf. A black couch is to the right of it .

MMT: This is a black keyboard. It is on the desk.
S2C: This is a square pillow. It is on the couch.
GT: This is a small square gray pillow. It is located on a black couch.

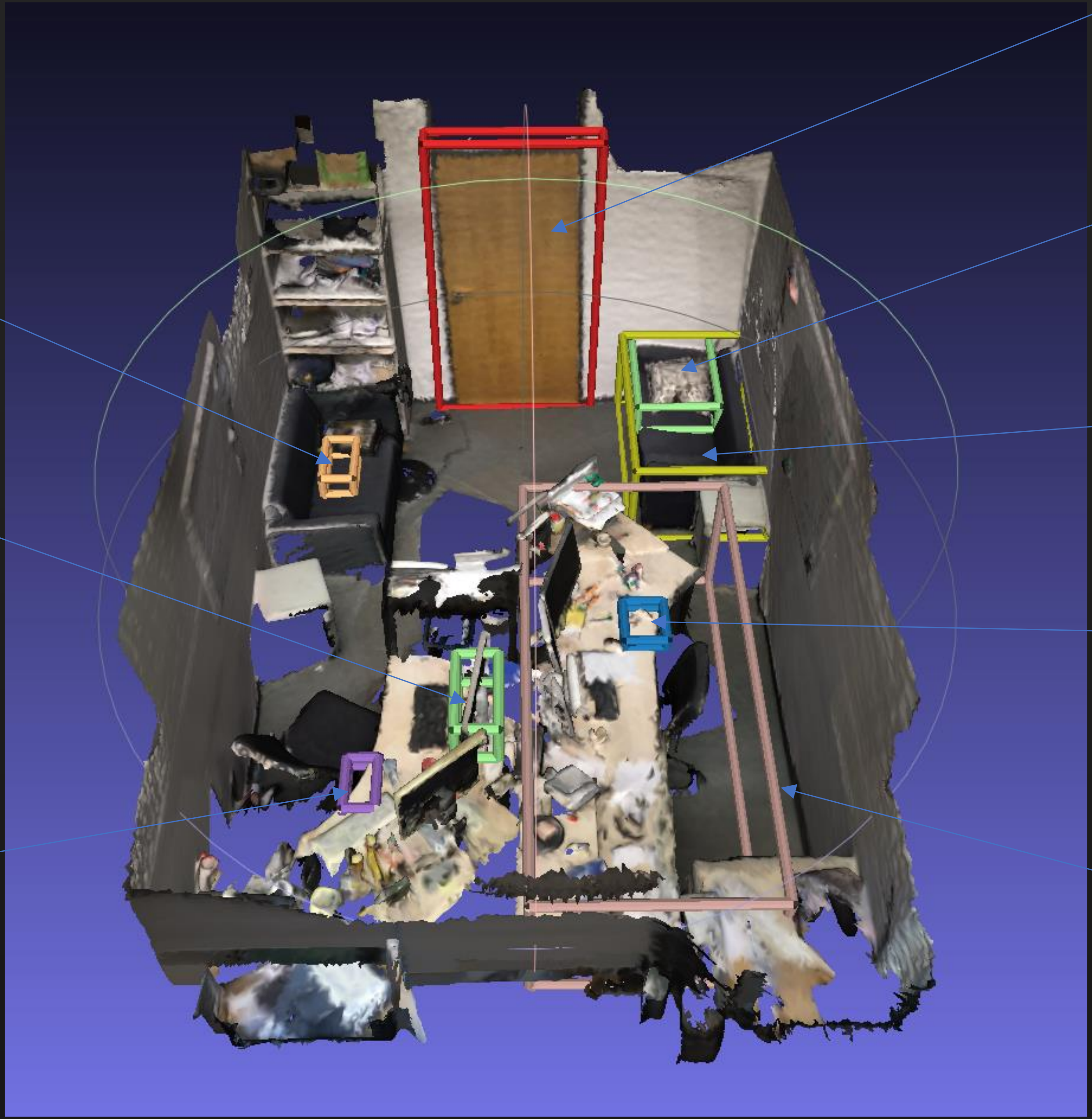
MMT: This is a black chair. It is to the right of the desk.
S2C: This is a black office chair. It is in front of a desk.
GT: This is a long tan desk. It is located near a wall and a small cabinet.

III. Qualitative Results

MMT: This is a black chair. It is to the right of the desk.
S2C: This is a brown couch. It is to the left of a brown table.
GT: It is a black sofa. It is located to the wall behind the fan.

MMT: This is a black monitor. It is on the desk.
S2C: N/A
GT: The monitor is located on top of the desk, and to the left of the other monitor facing the chair.

MMT: This is a black keyboard. It is on the desk.
S2C: N/A
GT: This is a long tan desk. It is located next to a black office chair.



MMT: This is a brown door. It is to the right of a bookshelf.
S2C: This is a white door. It is to the left of a couch.
GT: A light brown door beside a tall shelf. A black couch is to the right of it .

MMT: This is a black keyboard. It is on the desk.
S2C: This is a square pillow. It is on the couch.
GT: This is a small square gray pillow. It is located on a black couch.

MMT: This is a brown couch. It is to the right of the desk.
S2C: This is a brown couch. It is to the left of a table.
GT: The couch is located in the corner of the room. It is to the right side of the door.

2.MMT: This is a black keyboard. It is on a desk.
S2C: This is a black monitor. It is on a desk.
GT: A black computer screen is sitting on the desk. It is next to a black framed computer screen and to the left of it.

MMT: This is a black chair. It is to the right of the desk.
S2C: This is a black office chair. It is in front of a desk.
GT: This is a long tan desk. It is located near a wall and a small cabinet.

Scan2CapMMT

I. Scan2CapMMT Recap

II. Improving Scan2CapMMT

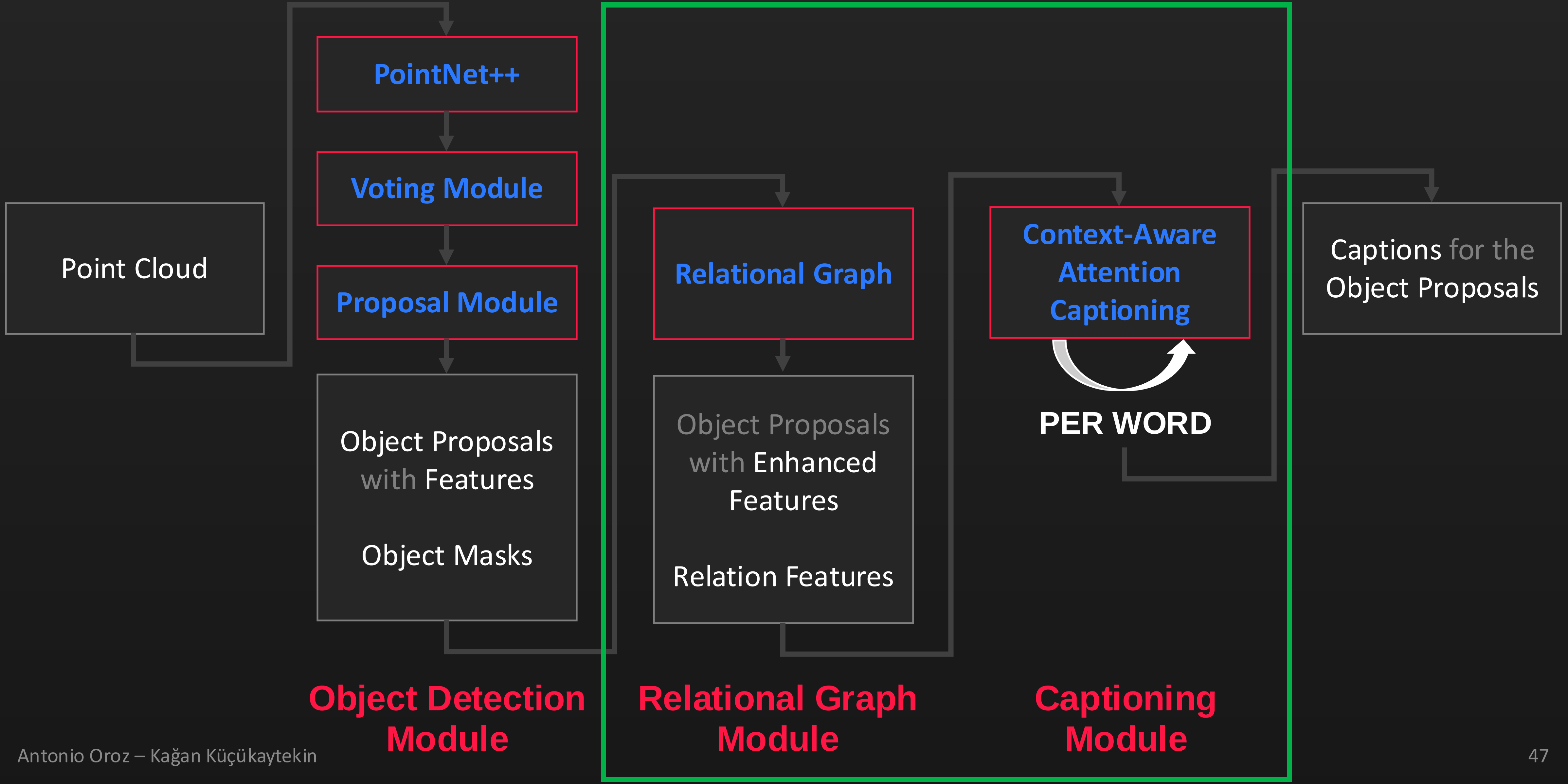
III. Quantitative & Qualitative Results

IV. Detection with Transformers

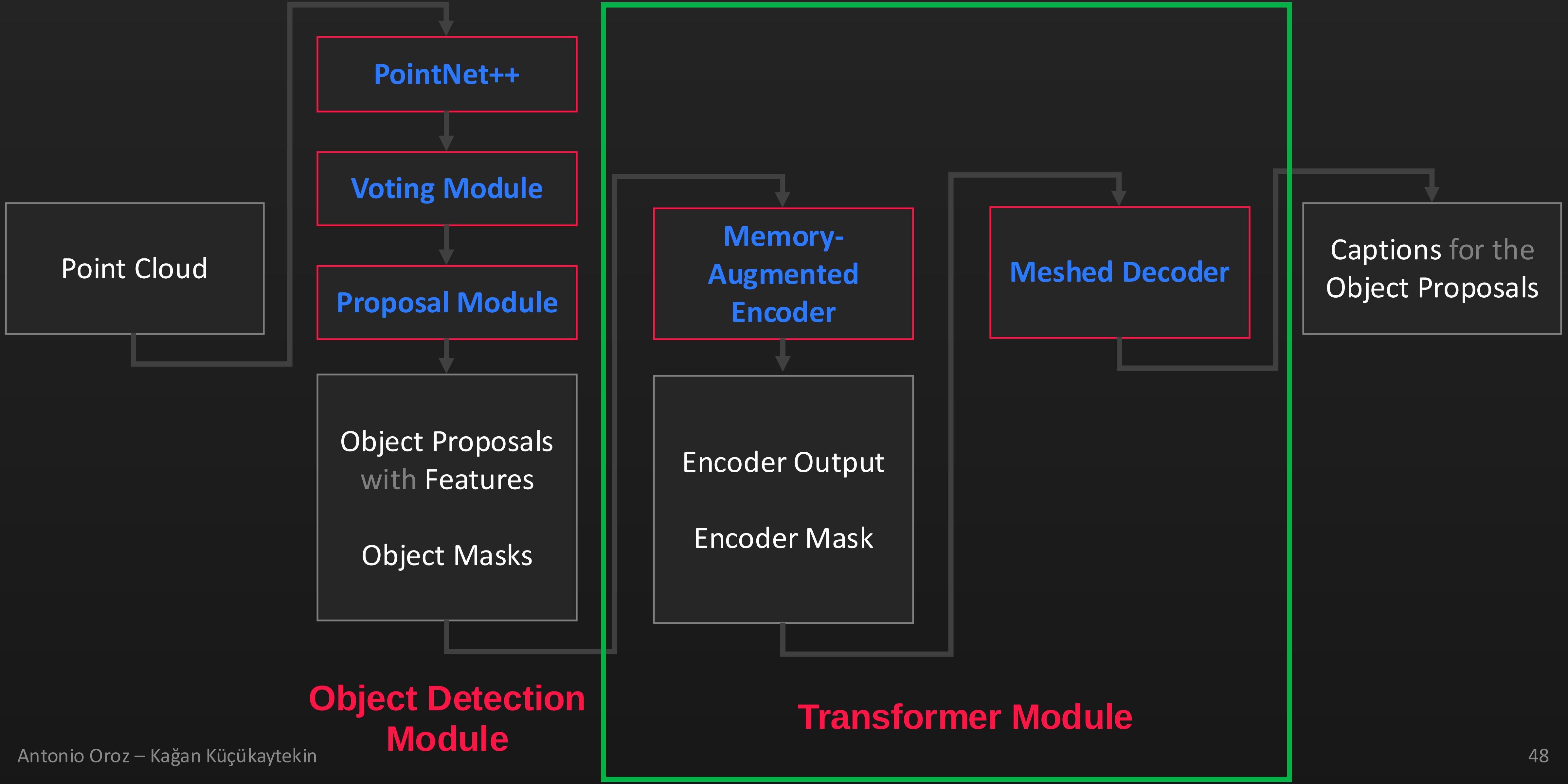
V. Timeline until the Final Presentation

Scan2CapMMT

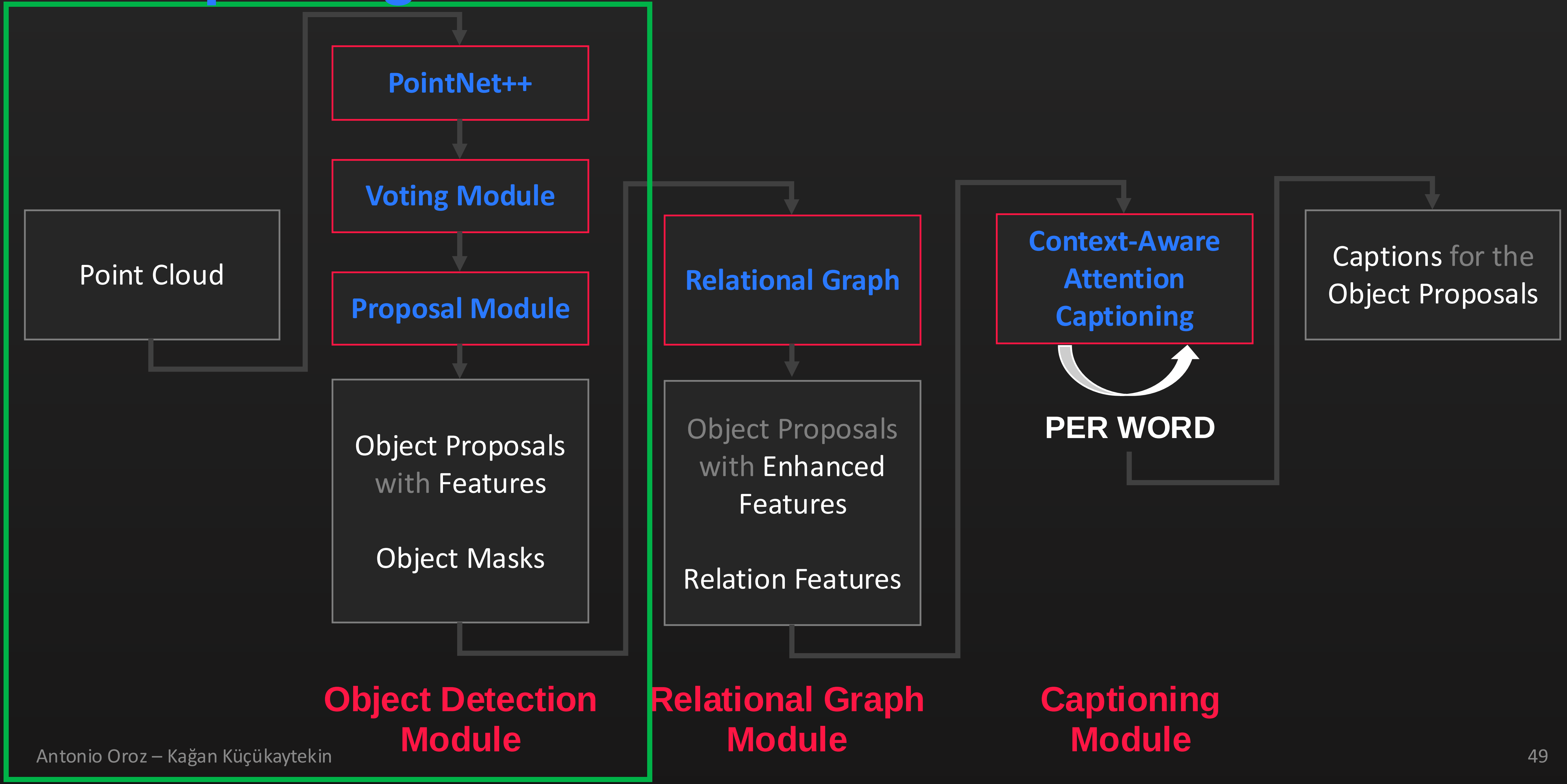
IV. Our Contribution



IV. Our Contribution



IV. Exploring Transformers for Detection Module



IV. Detection with Transformers

VoteNet

Point Cloud

PointNet++

Voting Module

Proposal Module

IV. Detection with Transformers

VoteNet

Point Cloud

↓

PointNet++

↓

Voting Module

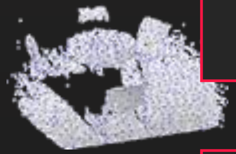
↓

Proposal Module

IV. Detection with Transformers

VoteNet

Point Cloud



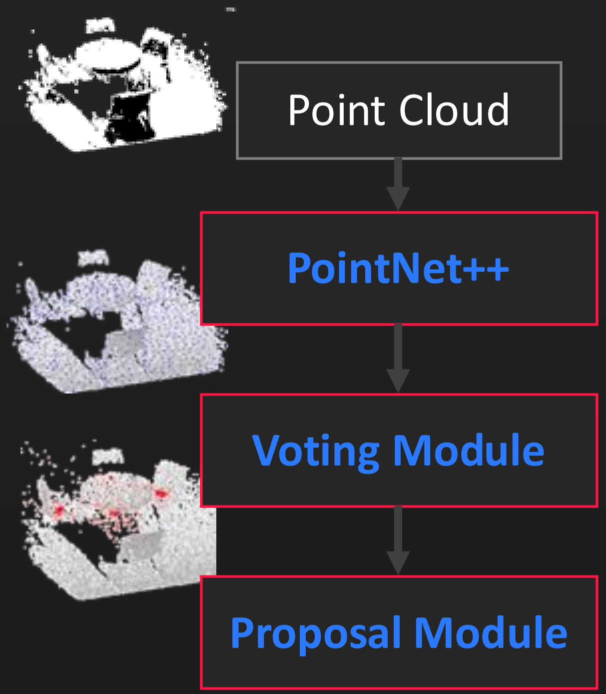
PointNet++

Voting Module

Proposal Module

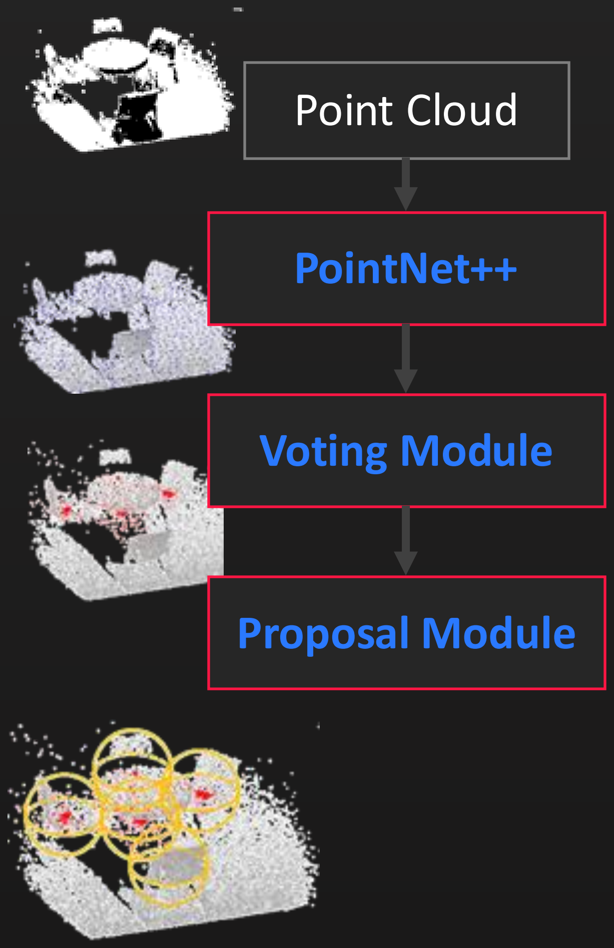
IV. Detection with Transformers

VoteNet



IV. Detection with Transformers

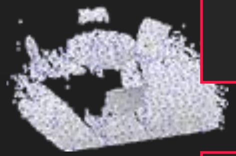
VoteNet



IV. Detection with Transformers

VoteNet

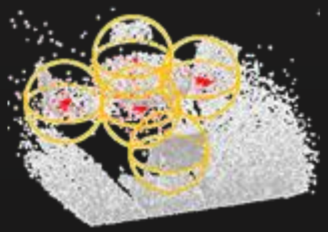
Point Cloud

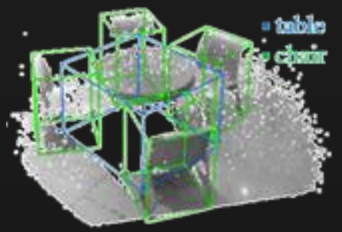


PointNet++

Voting Module

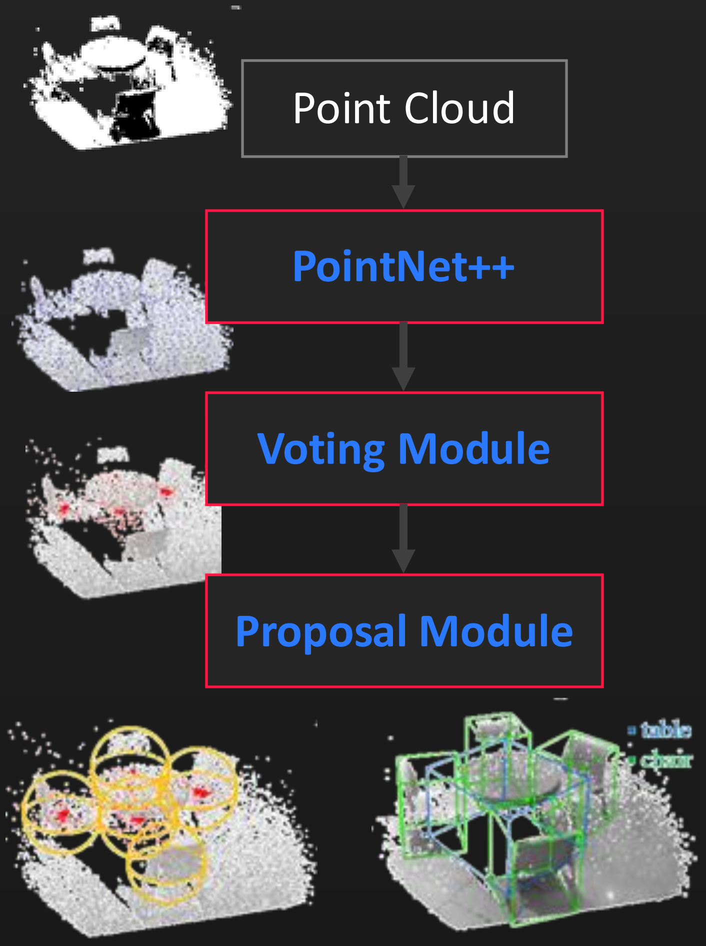
Proposal Module





IV. Detection with Transformers

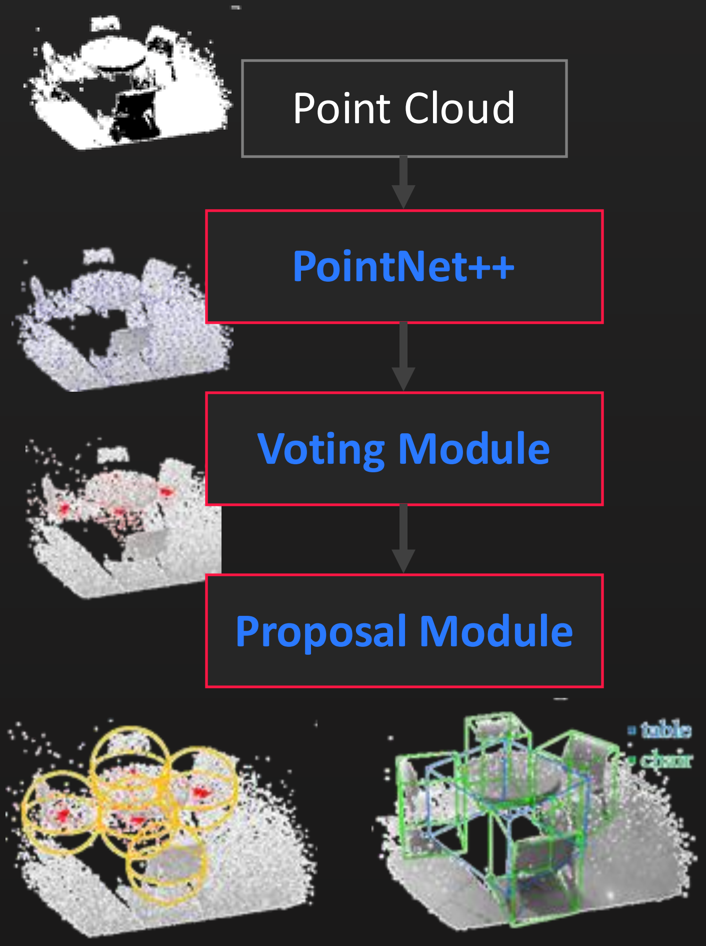
VoteNet



*Vote grouping is an issue! Especially when objects are overlapping.

IV. Detection with Transformers

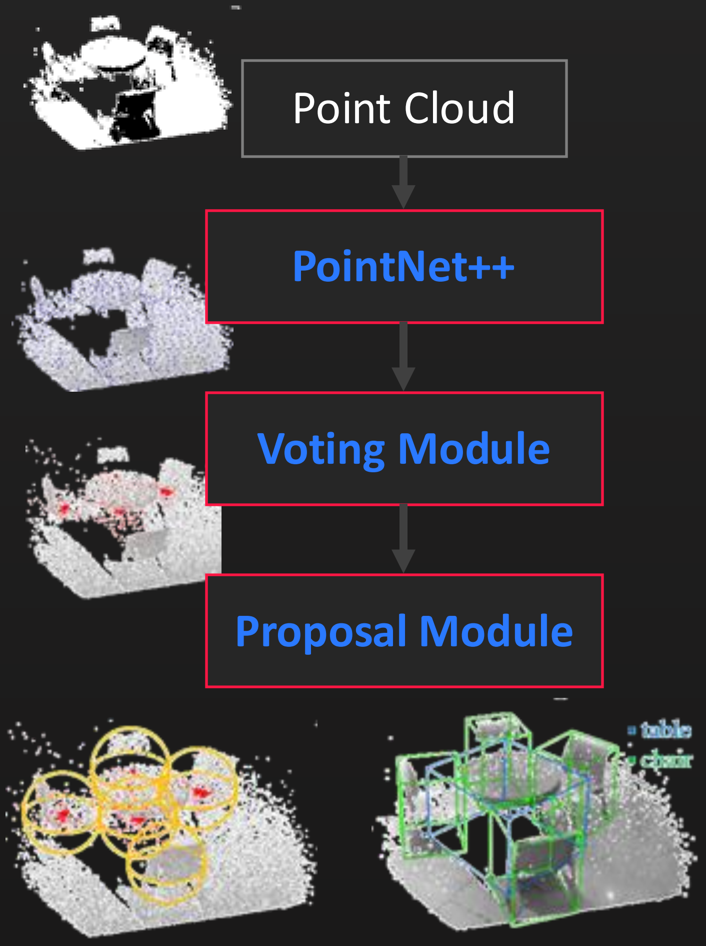
VoteNet



*Vote grouping is an issue! Especially when objects are overlapping.
*Radius for grouping is an important hyperparameter

IV. Detection with Transformers

VoteNet

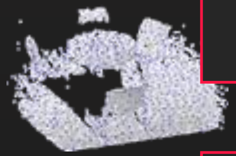


*Vote grouping is an issue! Especially when objects are overlapping.
*Radius for grouping is an important hyperparameter
*NMS only for eval

IV. Detection with Transformers

VoteNet

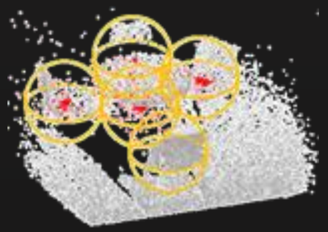
Point Cloud

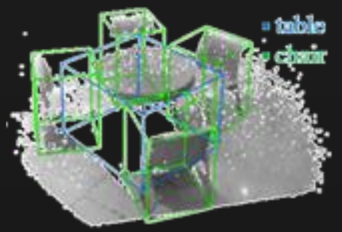


PointNet++

Voting

Proposal





3DETR

Point Cloud

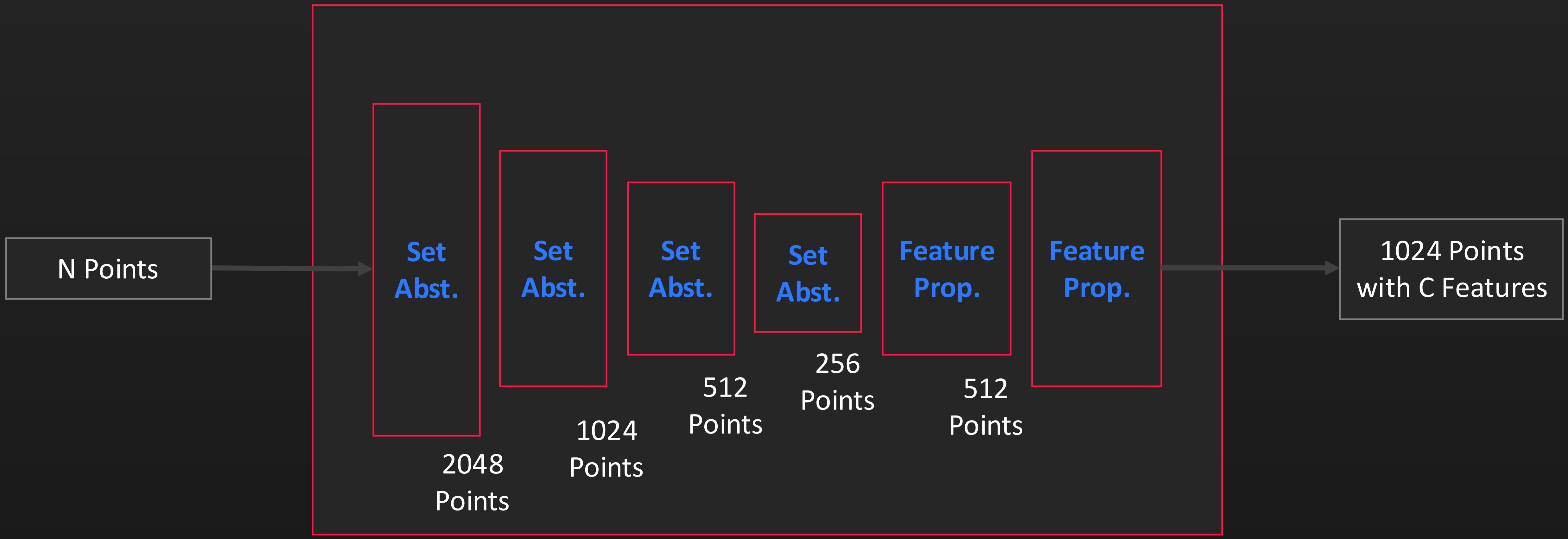
Group-Free-3D

Point Cloud

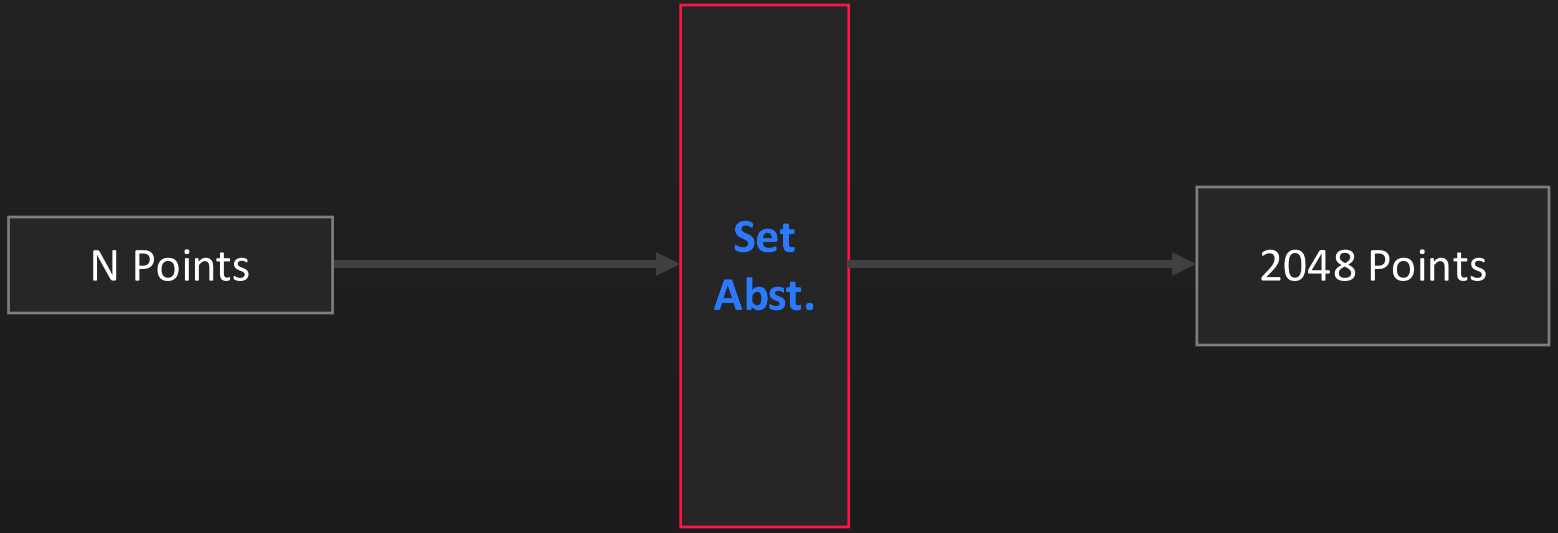
IV. Detection with Transformers

PointNet++

IV. Detection with Transformers

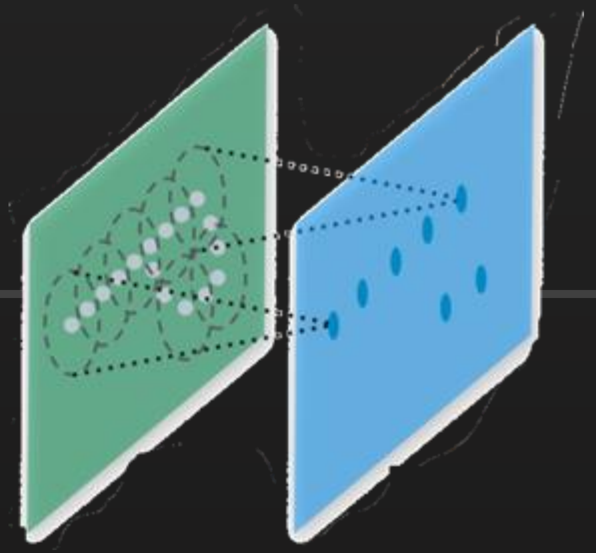


IV. Detection with Transformers



IV. Detection with Transformers

N Points

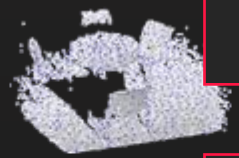


2048 Points

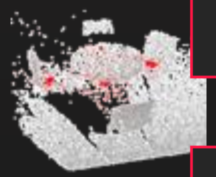
IV. Detection with Transformers

VoteNet

Point Cloud

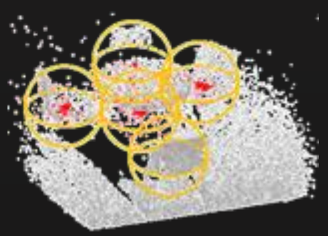


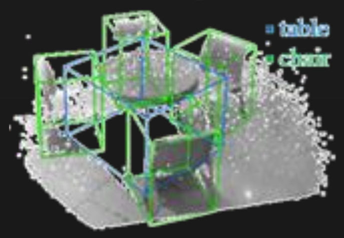
PointNet++



Voting

Proposal





3DETR

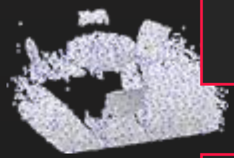
Point Cloud

Set Abstraction

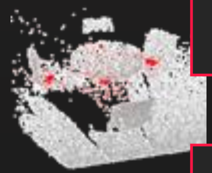
IV. Detection with Transformers

VoteNet

Point Cloud

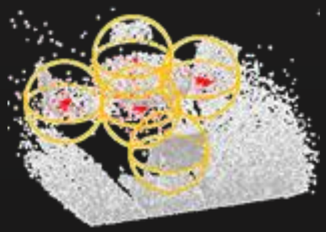


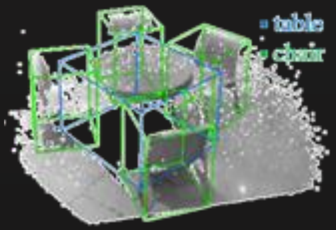
PointNet++



Voting

Proposal





3DETR

Point Cloud

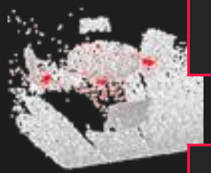
Set Abstraction

IV. Detection with Transformers

VoteNet

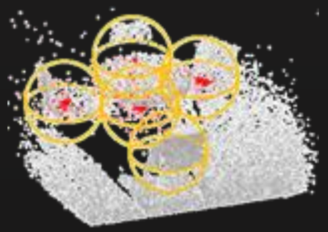
Point Cloud ↓

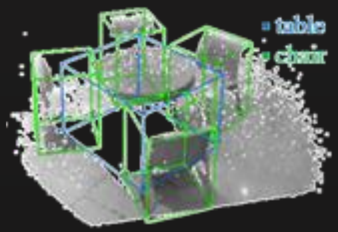
PointNet++ ↓



Voting ↓

Proposal





3DETR

Point Cloud ↓

Set Abstraction

Group-Free-3D

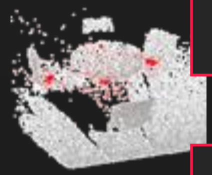
Point Cloud ↓

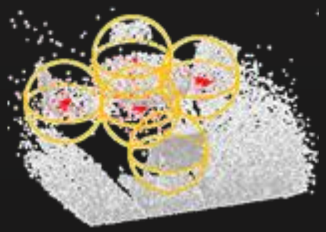
Pointnet++

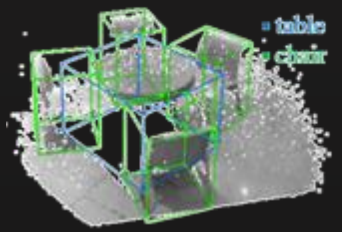
IV. Detection with Transformers

VoteNet

Point Cloud → **PointNet++** → **Voting** → **Proposal**







3DETR

Point Cloud → **Set Abstraction** → **Transformer** → **Proposal**

Group-Free-3D

Point Cloud → **Pointnet++** → **Transformer** → **Proposal**

IV. Detection with Transformers

VoteNet

3DETR

Group-Free-3D

- Vote grouping is an issue!
- Cluster radius
- NMS reliance

IV. Detection with Transformers

VoteNet

- Vote grouping is an issue!
- Cluster radius
- NMS reliance

3DETR

+No NMS

Group-Free-3D

+No NMS

IV. Detection with Transformers

VoteNet

- Vote grouping is an issue!
- Cluster radius
- NMS reliance

3DETR

+No NMS
+Predict with every decoder output

Group-Free-3D

+No NMS
+Predict with every decoder output

IV. Detection with Transformers

VoteNet

- Vote grouping is an issue!
- Cluster radius
- NMS reliance

3DETR

+No NMS
+Predict with every decoder output

Group-Free-3D

+No NMS
+Predict with every decoder output
+Use learnable pos. embeddings

IV. Detection with Transformers

VoteNet	3DETR	Group-Free-3D
- Vote grouping is an issue! - Cluster radius - NMS reliance	+No NMS +Predict with every decoder output	+No NMS +Predict with every decoder output +Use learnable pos. embeddings +More efficient point sampling strategy

IV. Detection with Transformers

VoteNet

- Vote grouping is an issue!
- Cluster radius
- NMS reliance

3DETR

+No NMS
+Predict with every decoder output
+Simplest, flexible

Group-Free-3D

+No NMS
+Predict with every decoder output
+Use learnable pos. embeddings
+More efficient point sampling strategy

IV. Detection with Transformers

VoteNet	3DETR	Group-Free-3D
mAP@0.5: 39.9	mAP@0.5: 47	mAP@0.5: 49

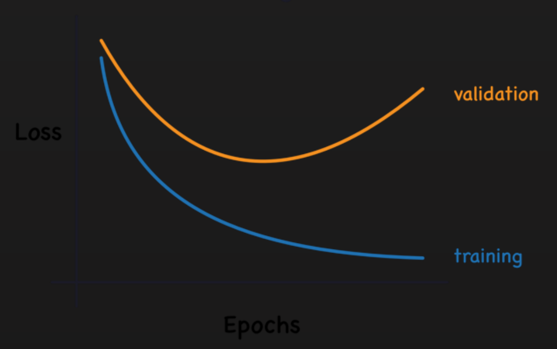
IV. Detection with Transformers

VoteNet	3DETR	Group-Free-3D
mAP@0.5: 39.9	mAP@0.5: 47	mAP@0.5: 49
1M Parameters	7.3M Parameters	14.5M Parameters

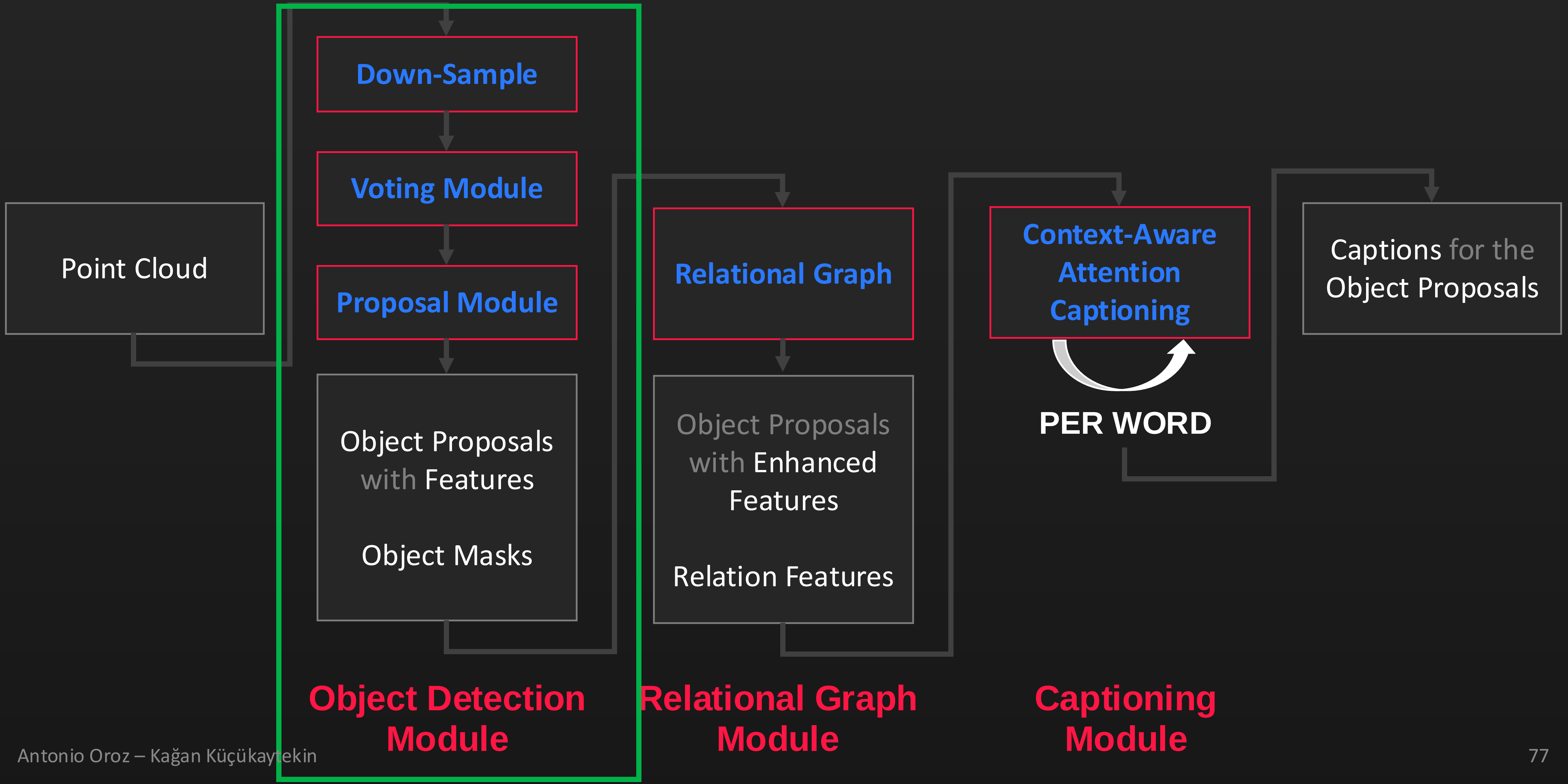
IV. Detection with Transformers

VoteNet	3DETR	Group-Free-3D
mAP@0.5: 39.9	mAP@0.5: 47	mAP@0.5: 49
1M Parameters	7.3M Parameters	14.5M Parameters

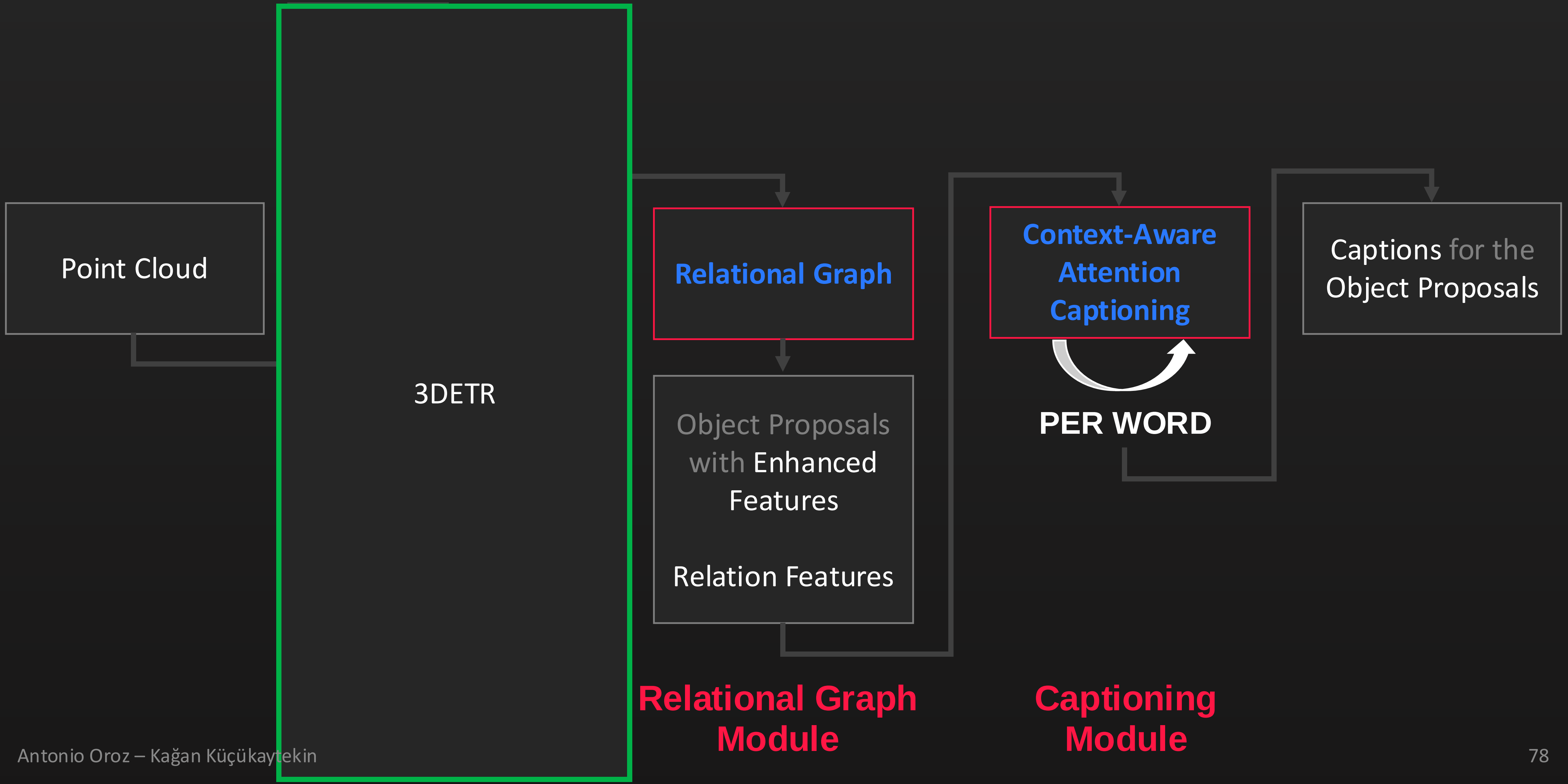
Because of data constraints, Group-Free-3D is more likely to overfit, so examine 3DETR



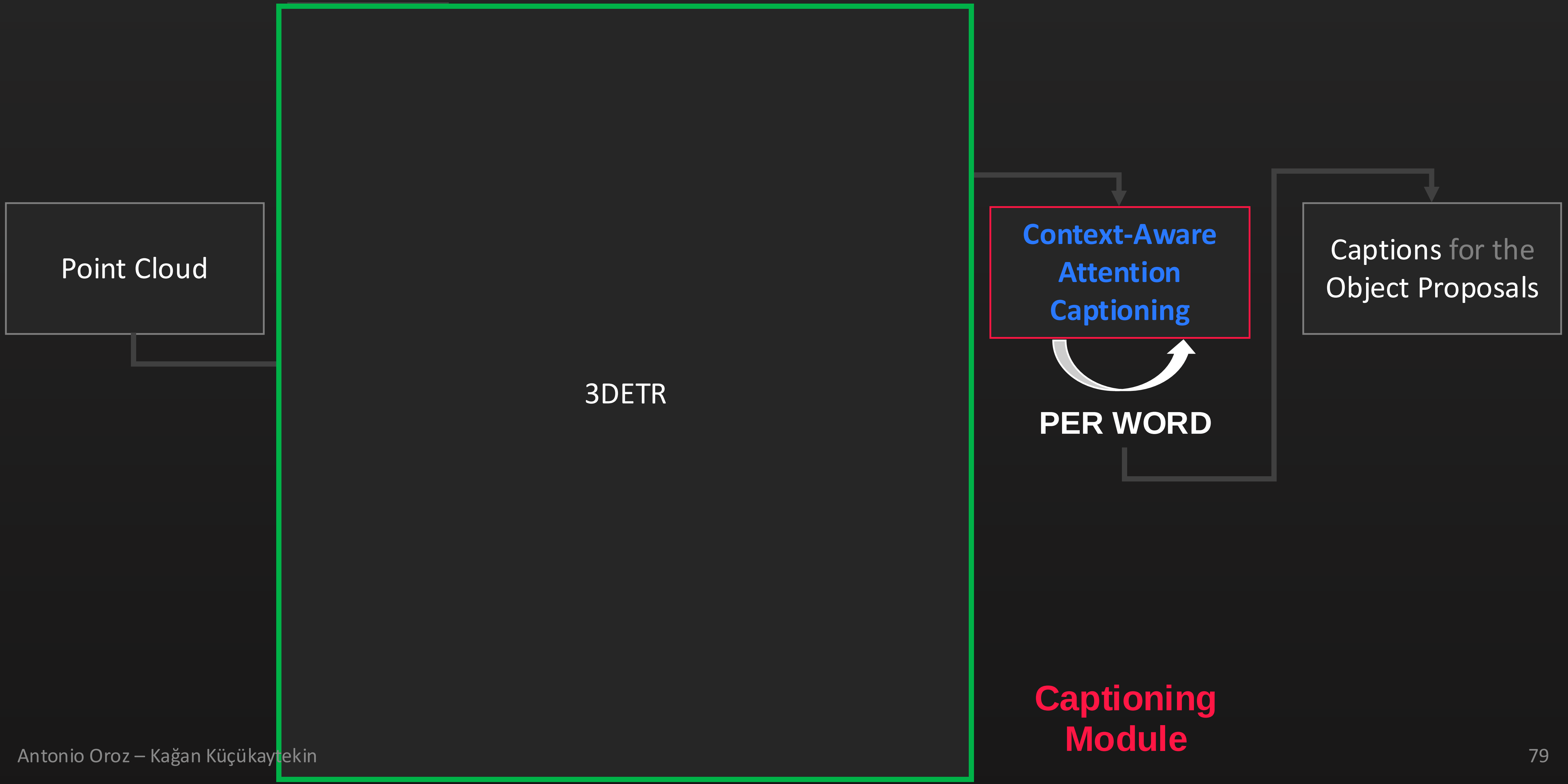
IV. Exploring Transformers for Detection Module



IV. Exploring Transformers for Detection Module



IV. Exploring Transformers for Detection Module



IV. Status

What we have done?

IV. Status

What we have done:

- Integrate 3DETR into the architecture

IV. Status

What we have done:

- Integrate 3DETR into the architecture
- Test end-to-end pipeline by overfitting to single sample without caption

IV. Status

What we have done:

- Integrate 3DETR into the architecture
- Test end-to-end pipeline by overfitting to single sample without caption

Possible next steps?

IV. Status

What we have done:

- Integrate 3DETR into the architecture
- Test end-to-end pipeline by overfitting to single sample without caption

Possible next steps:

- End-to-end overfit to small sample for whole task

IV. Status

What we have done:

- Integrate 3DETR into the architecture
- Test end-to-end pipeline by overfitting to single sample without caption

Possible next steps:

- End-to-end overfit to small sample for whole task
- Try transfer Learning with pre-trained 3DETR-m

IV. Status

What we have done:

- Integrate 3DETR into the architecture
- Test end-to-end pipeline by overfitting to single sample without caption

Possible next steps:

- End-to-end overfit to small sample for whole task
- Try transfer Learning with pre-trained 3DETR-m
- No promise! Ablation studies on our model is our Prio 1.

Scan2CapMMT

I. Scan2CapMMT Recap

II. Improving Scan2CapMMT

III. Quantitative & Qualitative Results

IV. Detection with Transformers

V. Timeline until the Final Presentation

Scan2CapMMT

I. Scan2CapMMT Recap

II. Improving Scan2CapMMT

III. Quantitative & Qualitative Results

IV. Detection with Transformers

V. Timeline until the Final Presentation

V. Timeline until Final Presentation

- Reproduced Scan2Cap
- Started MMT Implemtation

1. Presentation

V. Timeline until Final Presentation

- Reproduced Scan2Cap
- Started MMT Implemtation

1. Presentation

- Finalized MMT Implementation
- Implemented Beam Search and RL
- Looked into Detection Module alternatives
- Prototype 3DETR Implementation into Scan2Cap pipeline

2. Presentation

V. Timeline until Final Presentation

- Reproduced Scan2Cap
- Started MMT Implemtation

1. Presentation

- Finalized MMT Implementation
- Implemented Beam Search and RL
- Looked into Detection Module alternatives
- Prototype 3DETR Implementation into Scan2Cap pipeline

2. Presentation

- Trying 3DETR with MMT
- Figuring out why Reinforcement Learning is unstable
- Qualitative and Quantitative Analysis
- Ablation Study

Final Presentation

Scan2CapMMT

I. Scan2CapMMT Recap

II. Improving Scan2CapMMT

III. Quantitative & Qualitative Results

IV. Detection with Transformers

V. Timeline until the Final Presentation

Scan2CapMMT

I. Scan2CapMMT Recap

II. Improving Scan2CapMMT

III. Quantitative & Qualitative Results

IV. Detection with Transformers

V. Timeline until the Final Presentation

THANK YOU FOR YOUR ATTENTION :D

Scan2CapMMT

Dense Captioning for 3D Scenes with Transformers

Scan2CapMMT

I. Scan2Cap

II. Meshed-Memory Transformer

III. Scan2CapMMT

IV. Insights & First Results

V. Next Steps

Scan2CapMMT

I. Scan2Cap

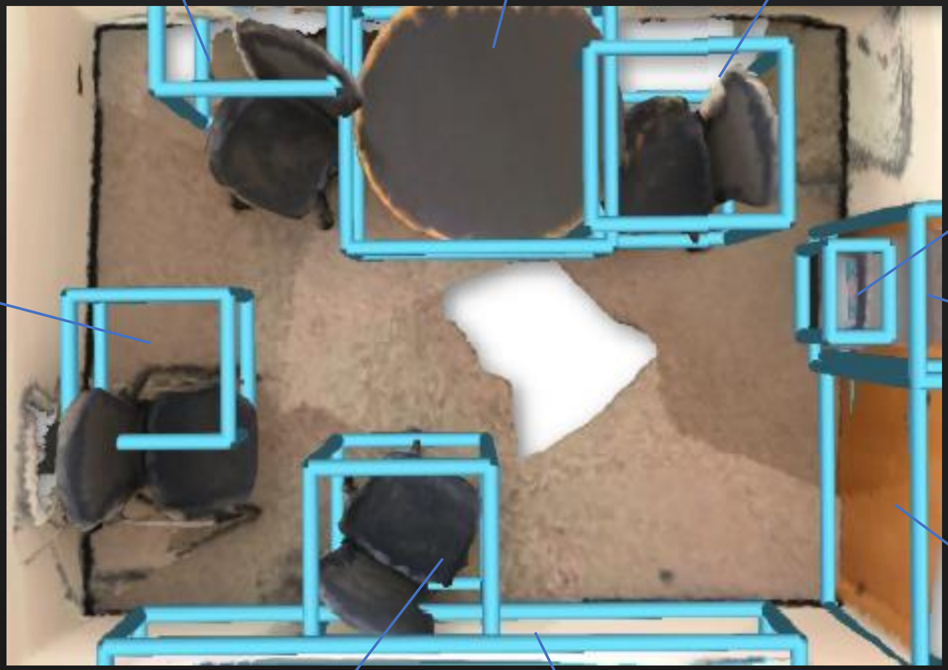
II. Meshed-Memory Transformer

III. Scan2CapMMT

IV. Insights & First Results

V. Next Steps

I. Scan2Cap: 3D Dense Captioning



I. Scan2Cap

Point Cloud → Captions for the Object Proposals

I. Scan2Cap: Architecture

Point Cloud

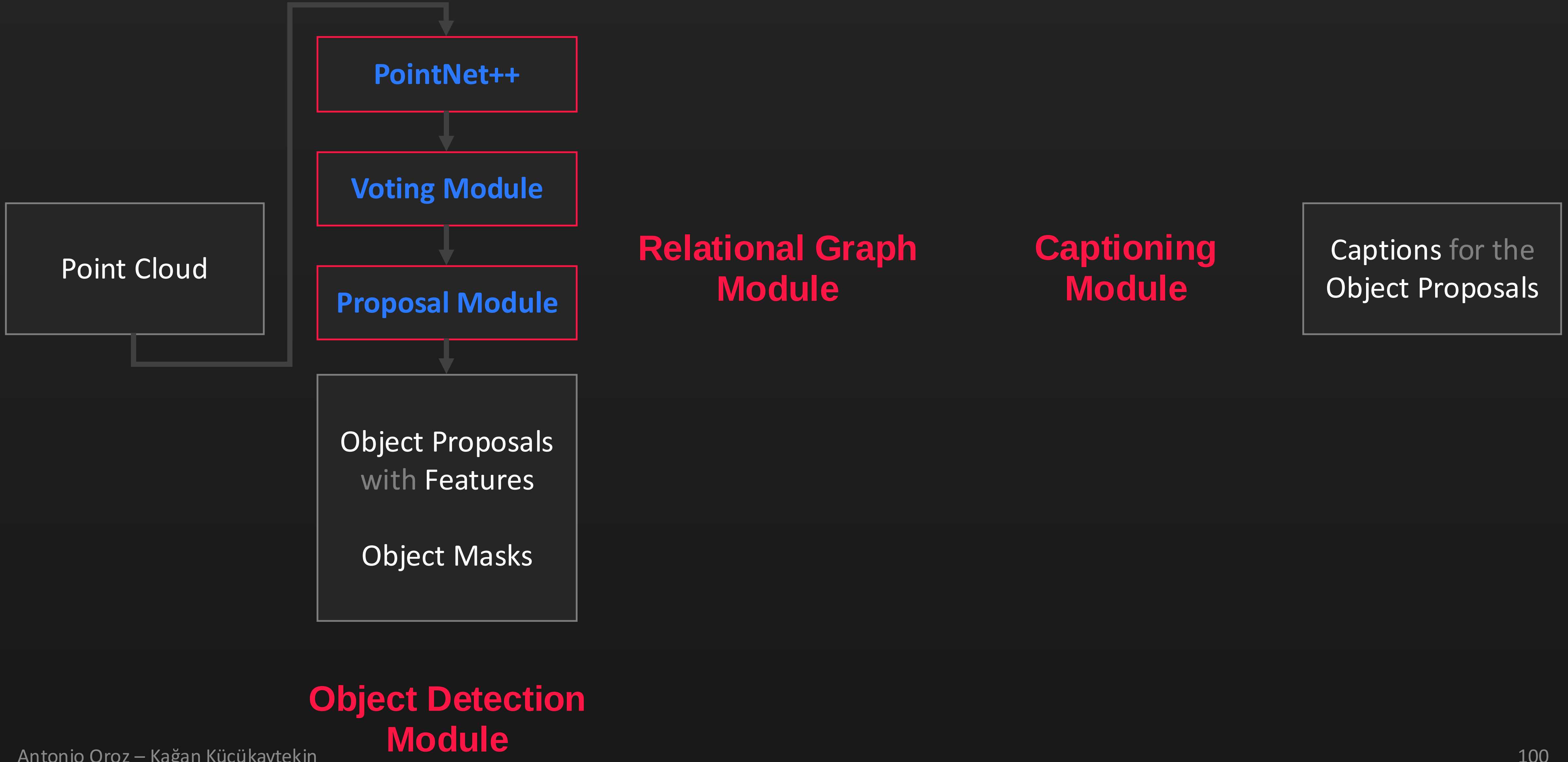
Object Detection Module

Relational Graph Module

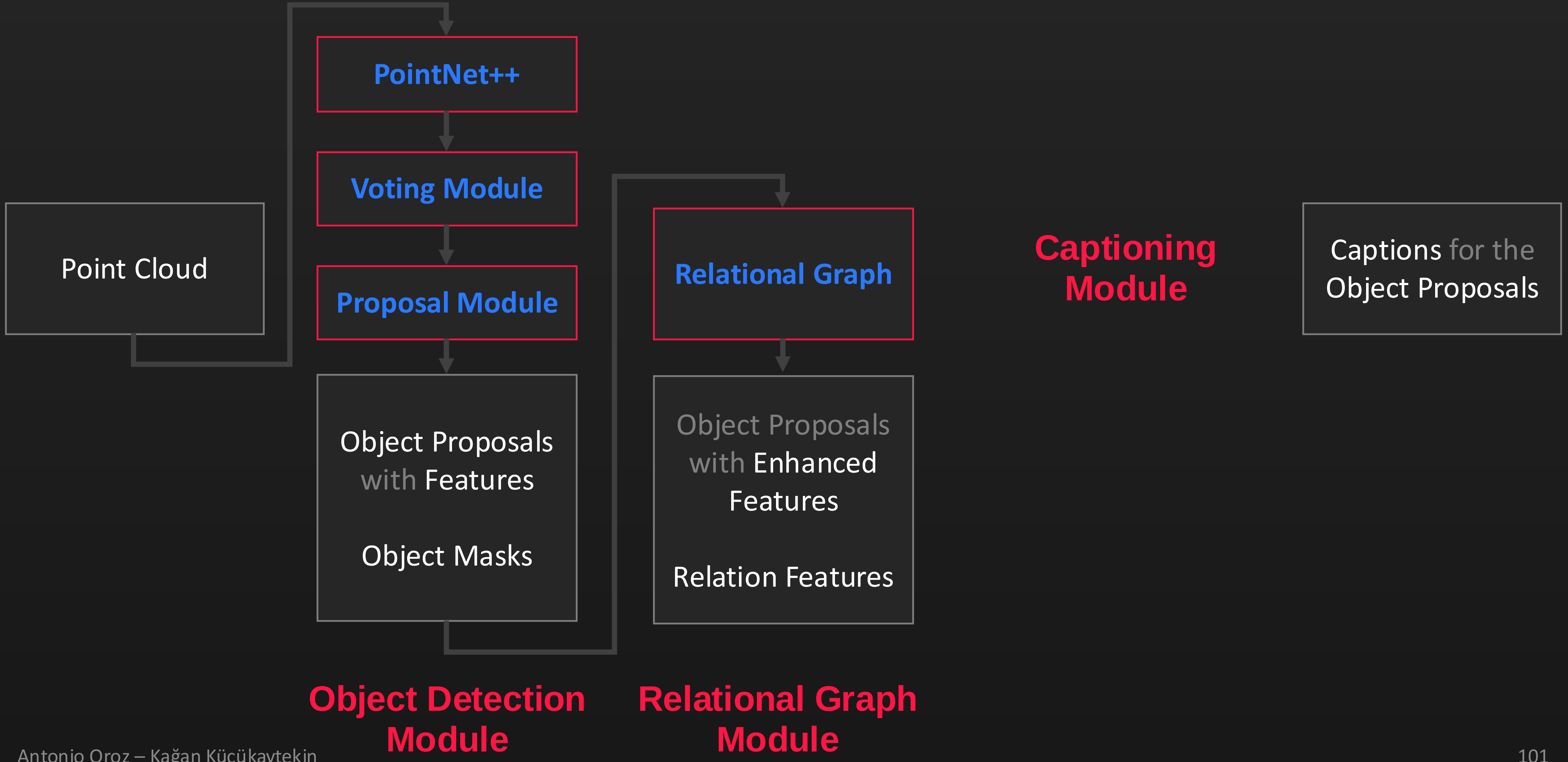
Captioning Module

Captions for the Object Proposals

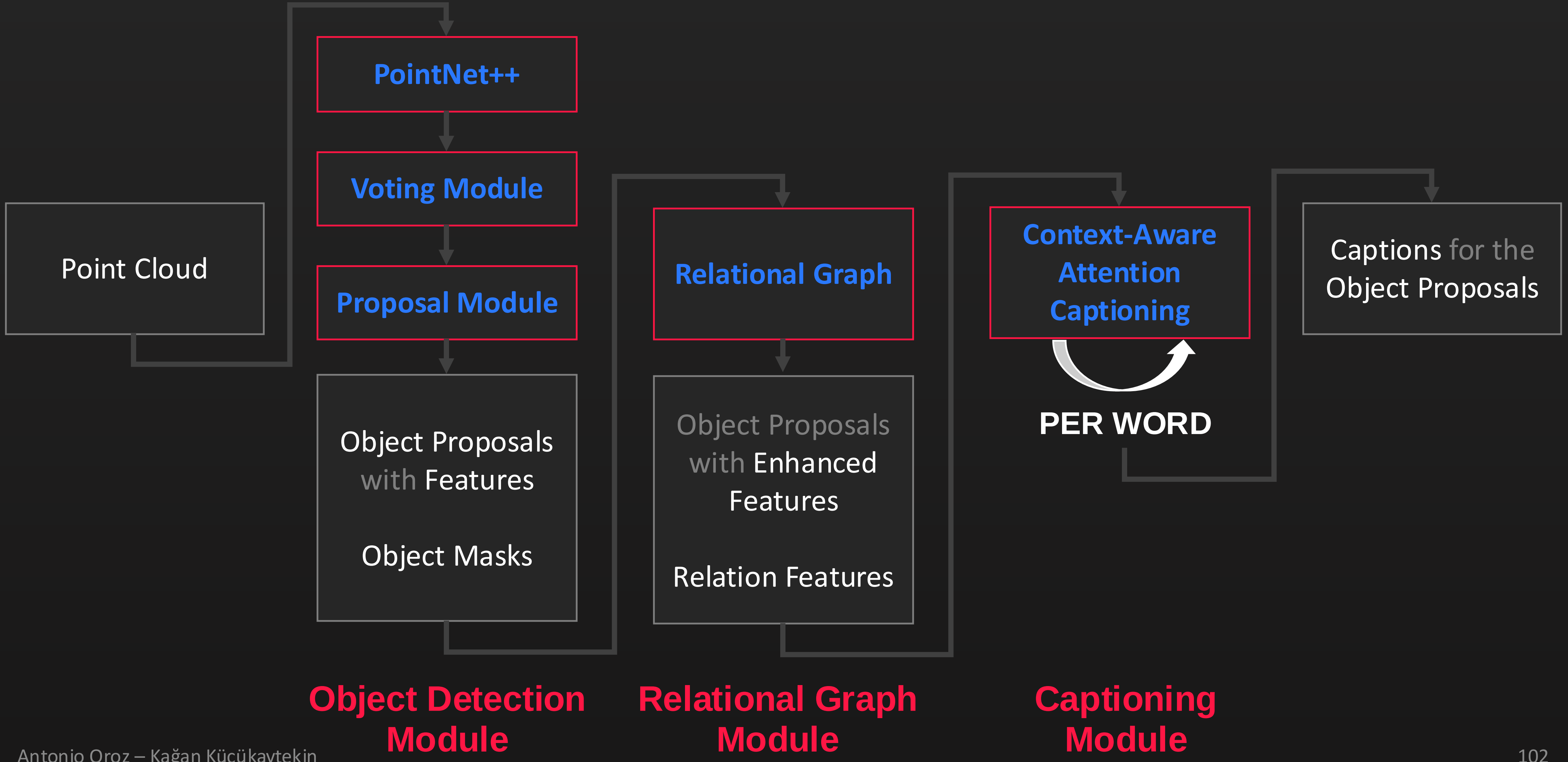
I. Scan2Cap: Architecture



I. Scan2Cap: Architecture



I. Scan2Cap: Architecture



Scan2CapMMT

I. Scan2Cap

II. Meshed-Memory Transformer

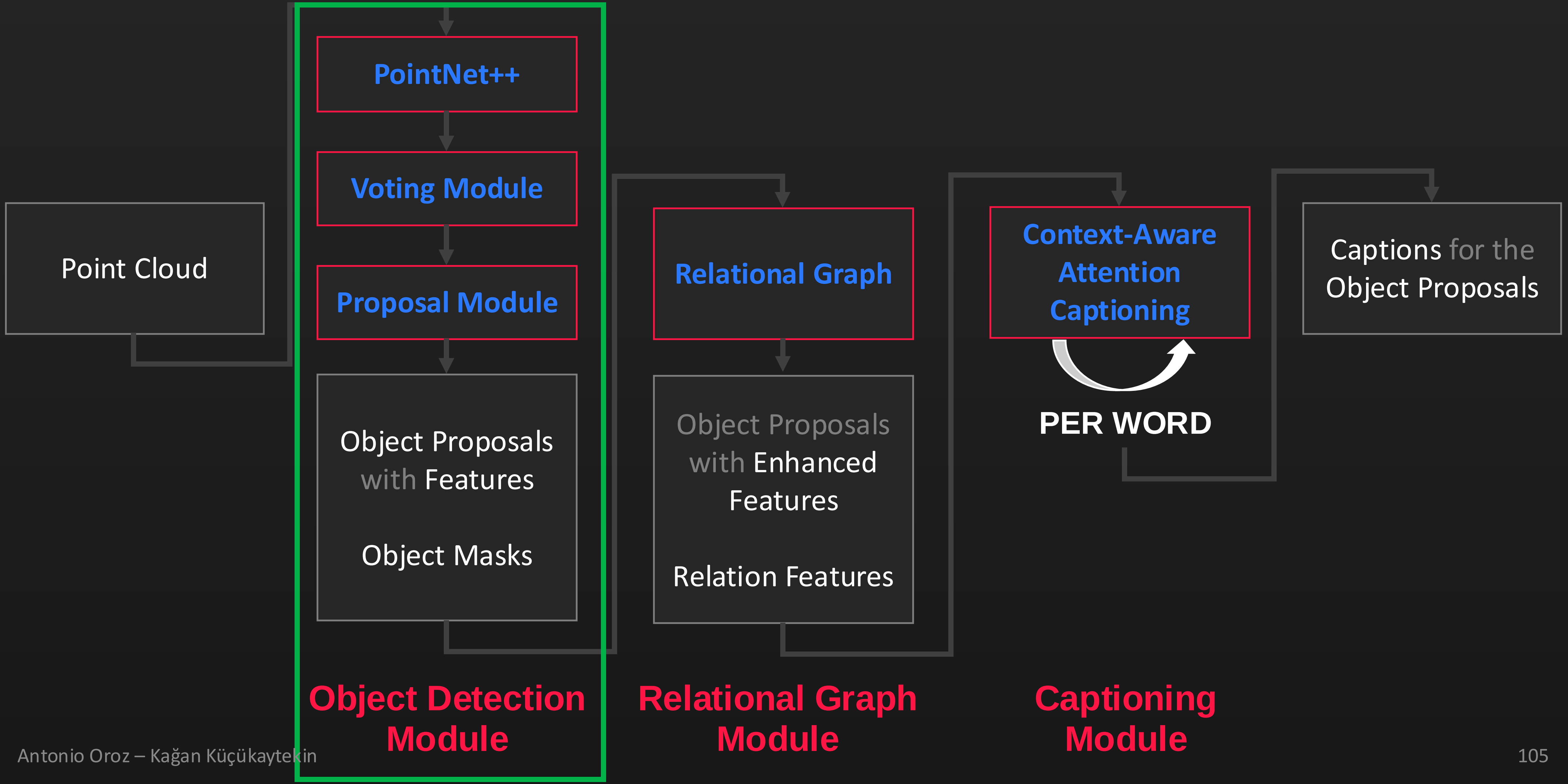
III. Scan2CapMMT

IV. Insights & First Results

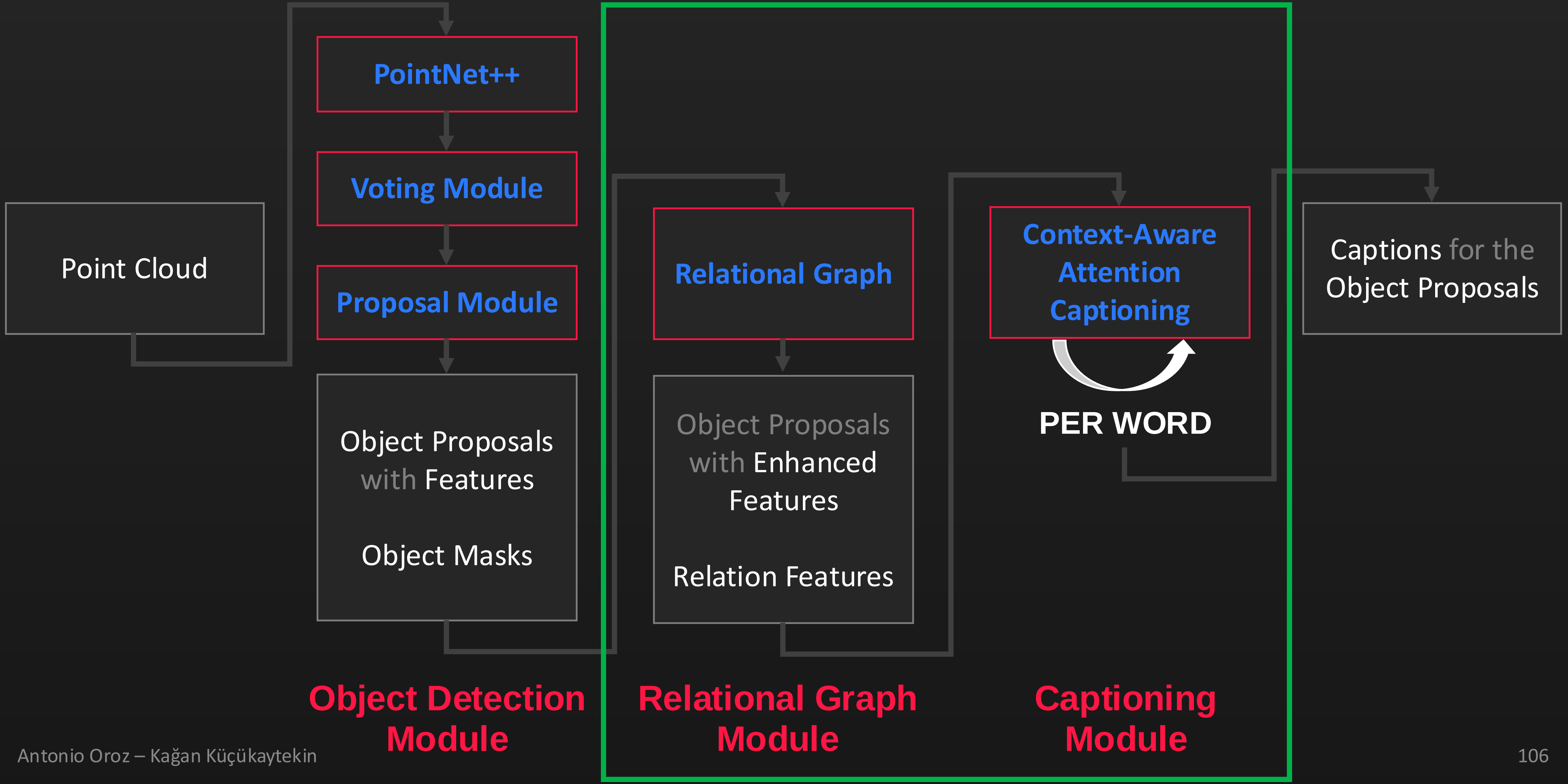
V. Next Steps

Scan2CapMMT

II. Scan2CapMMT: Motivation for MMT



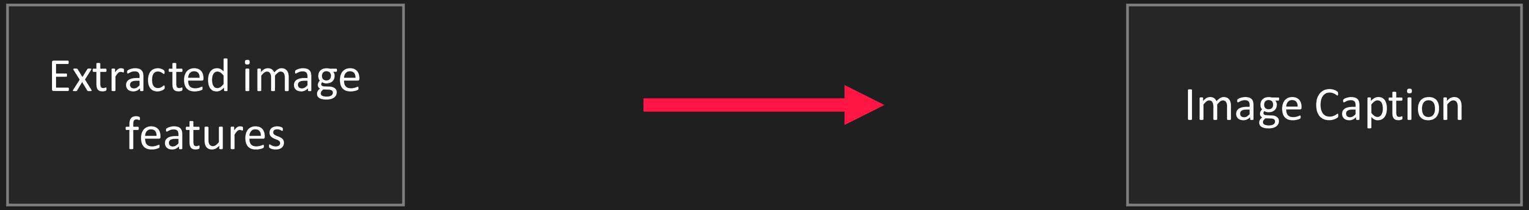
II. Scan2CapMMT: Motivation for MMT



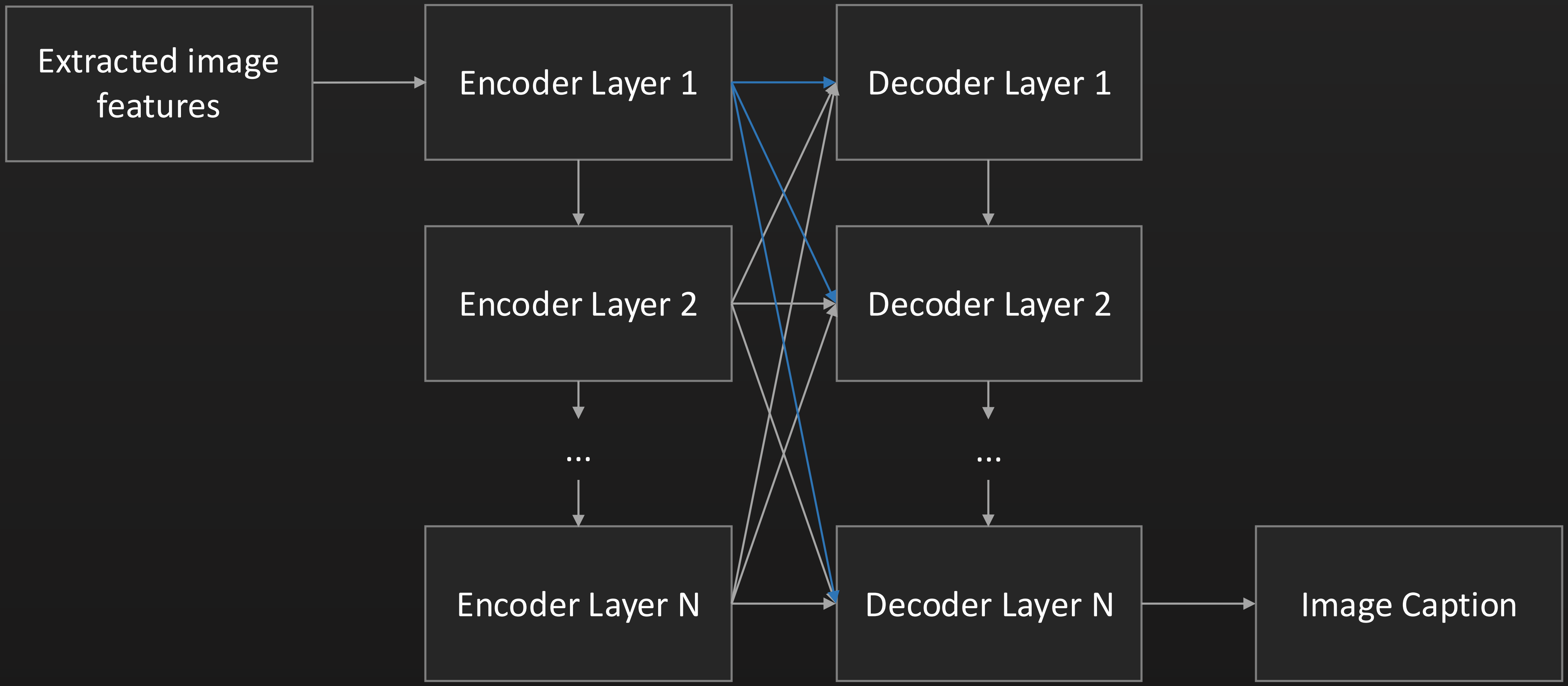
II. Scan2Cap

Point Cloud → Captions for the Object Proposals

II. Meshed-Memory Transformer

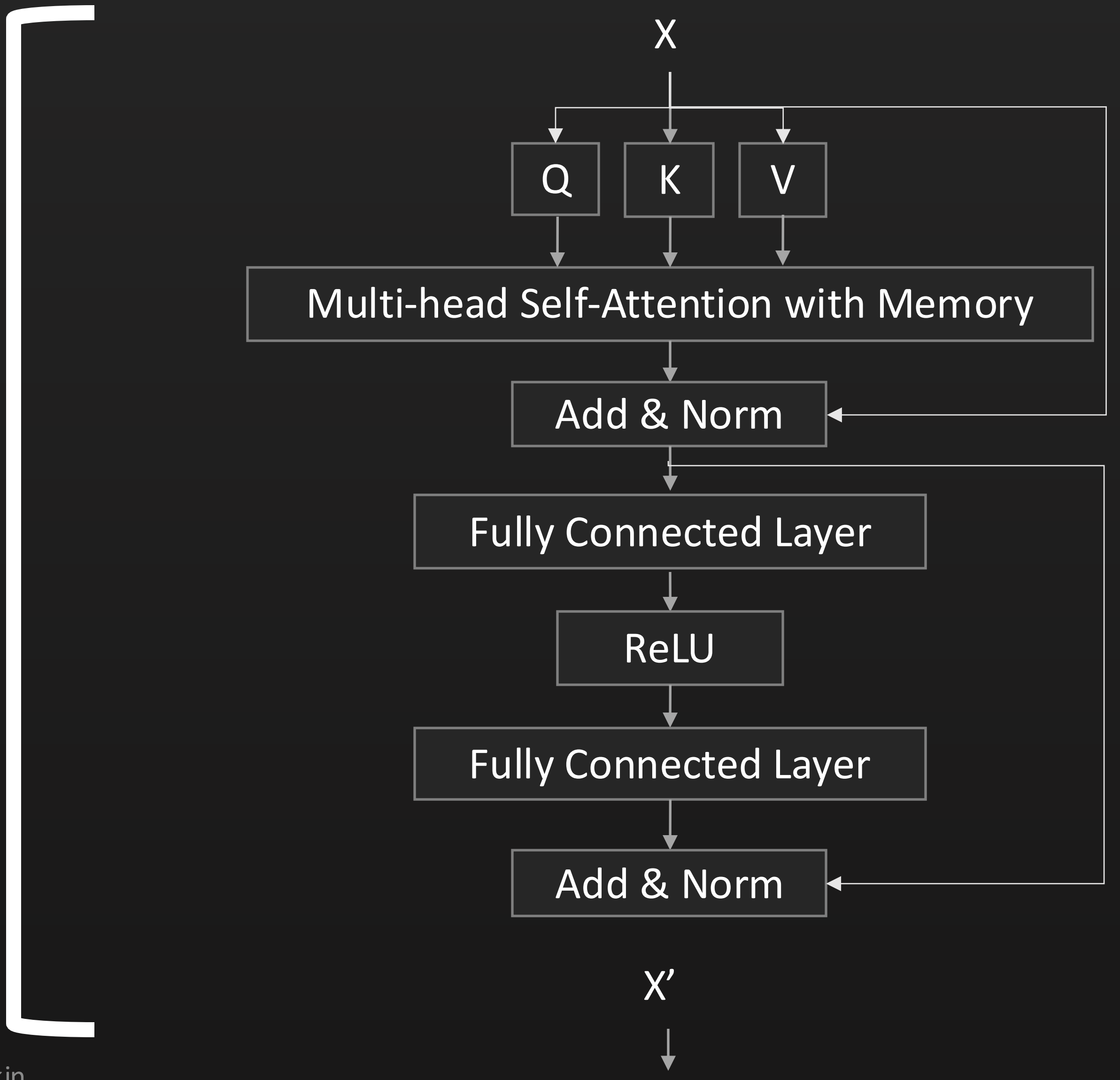


II. Meshed-Memory Transformer

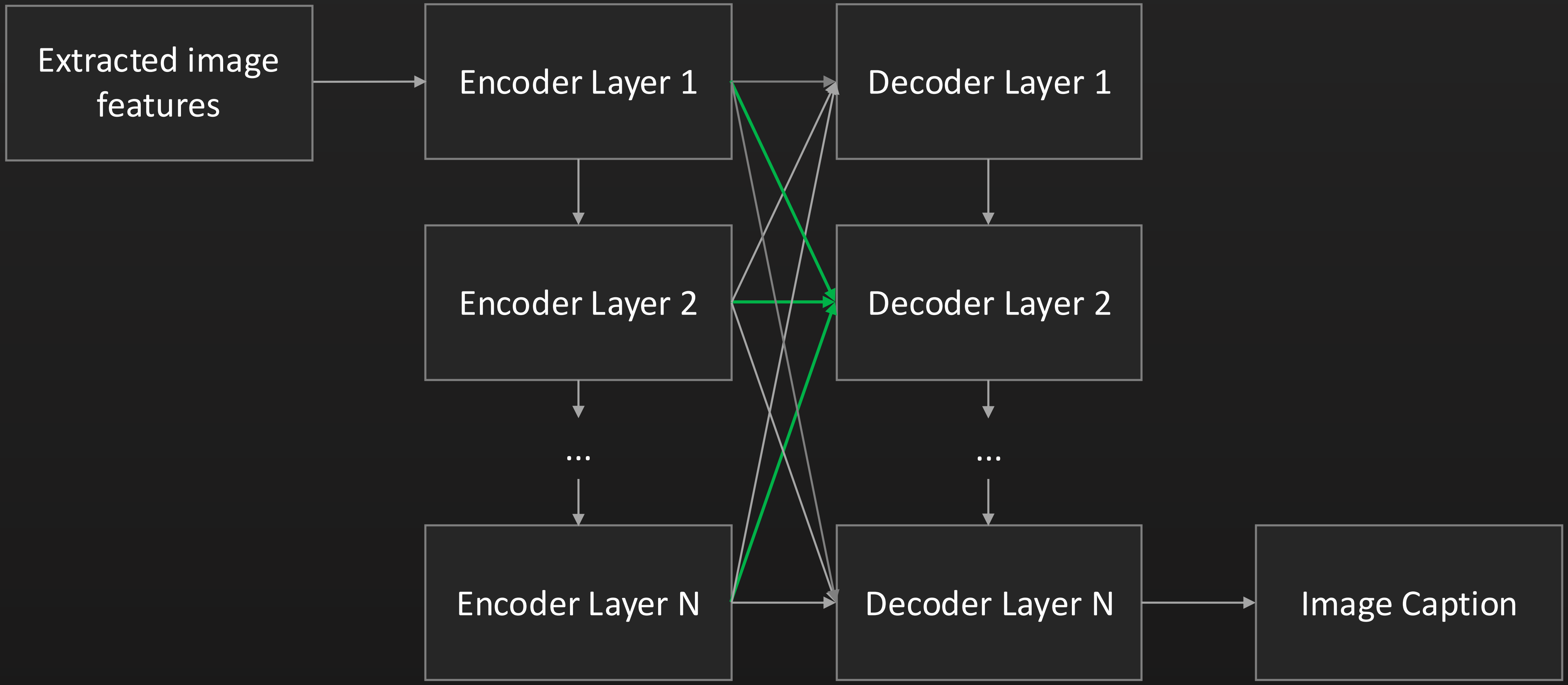


II. Meshed-Memory Transformer: Encoder

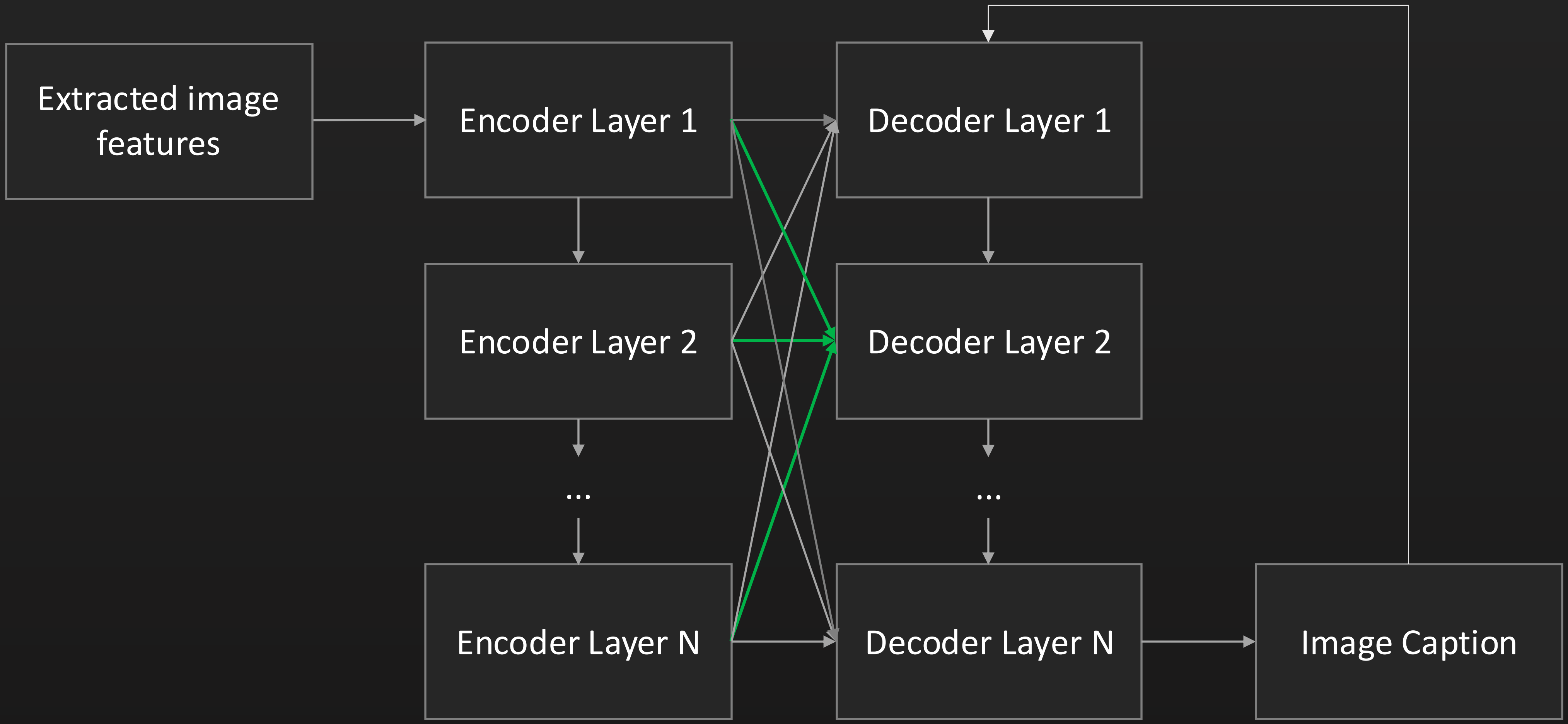
Encoder Layer 1



II. Meshed-Memory Transformer

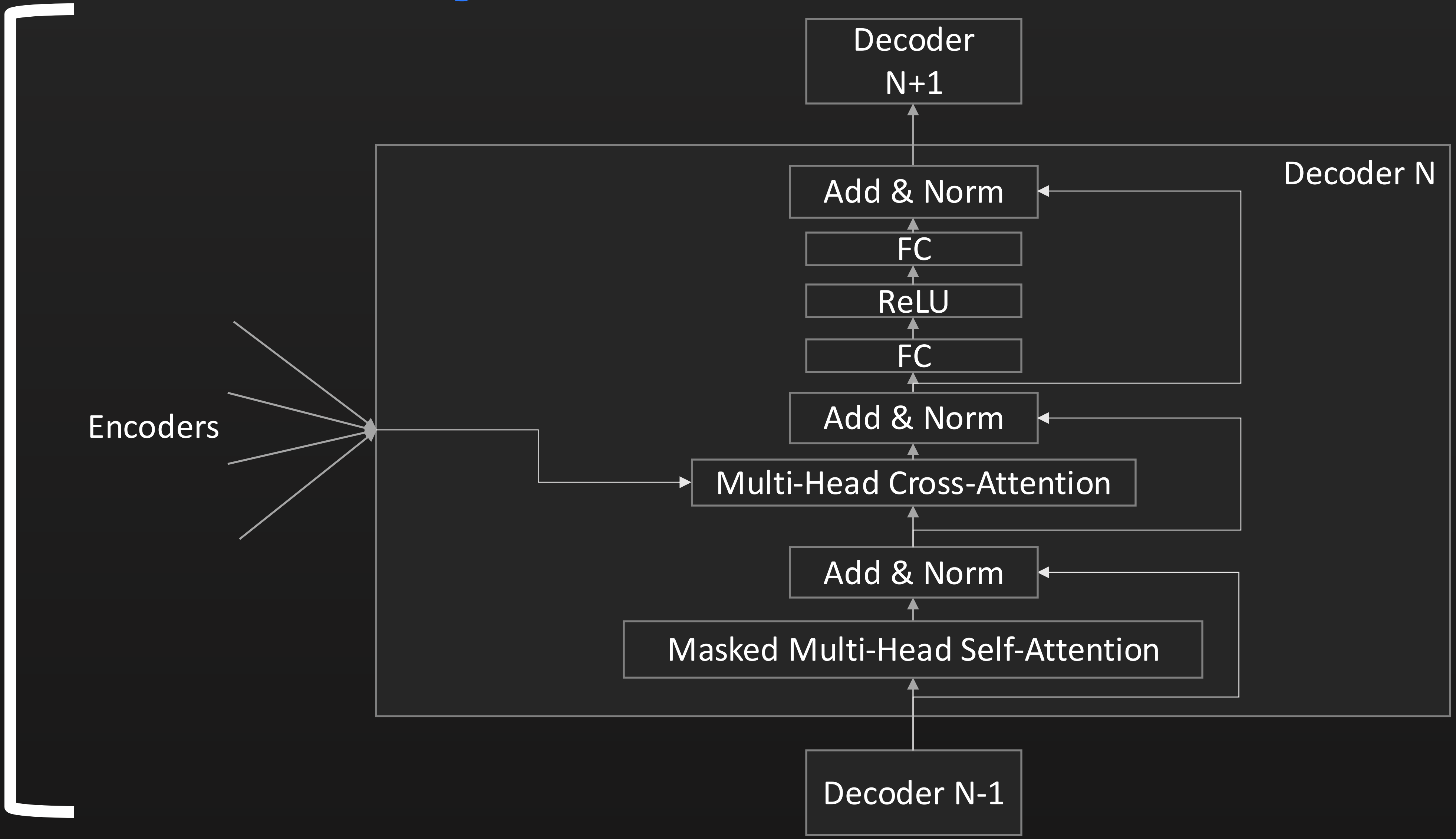


II. Meshed-Memory Transformer



II. Meshed-Memory Transformer: Decoder

Decoder Layer N



Scan2CapMMT

I. Scan2Cap

II. Meshed-Memory Transformer

III. Scan2CapMMT

IV. Insights & First Results

V. Next Steps

Scan2CapMMT

I. Scan2Cap

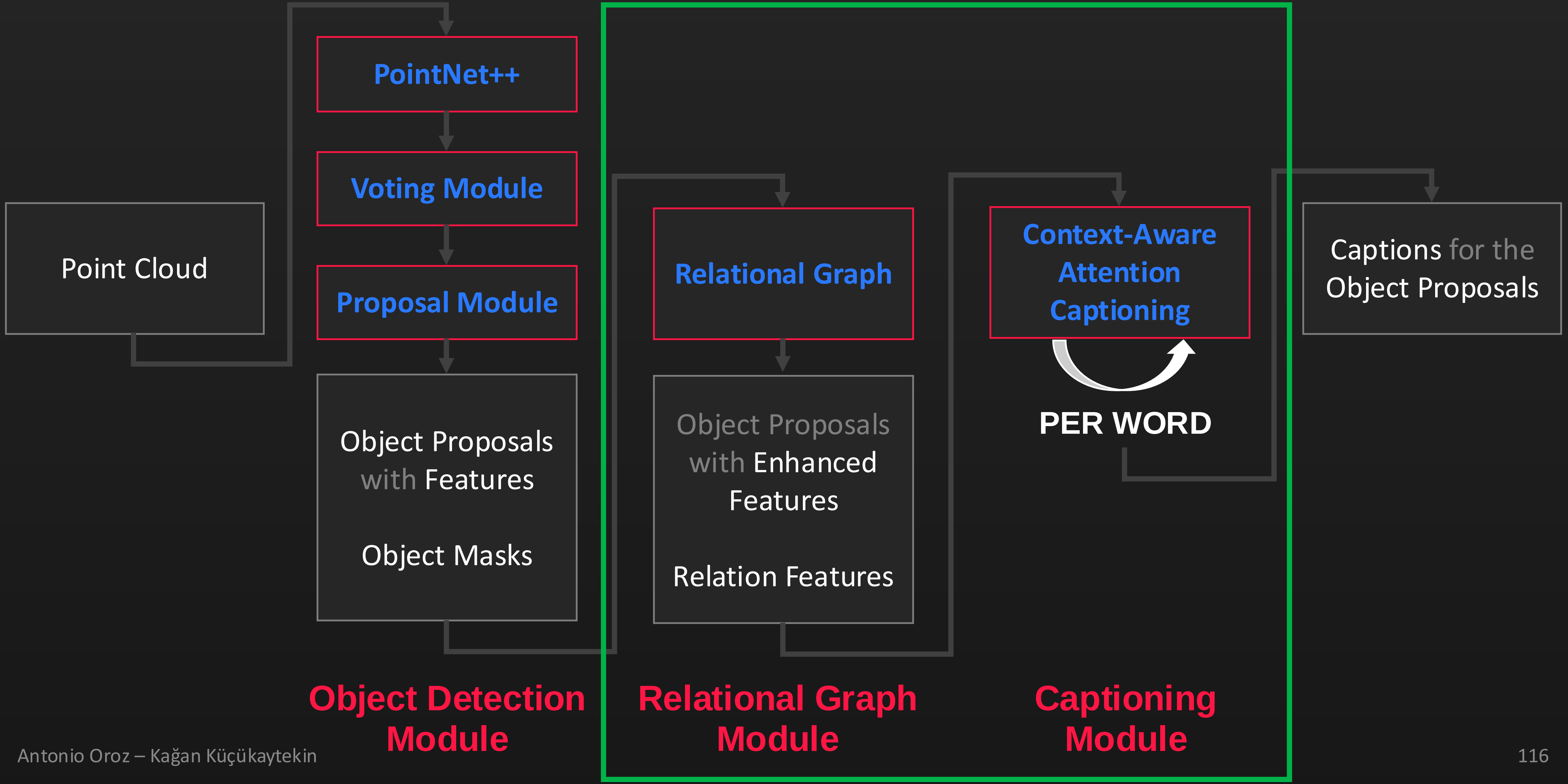
II. Meshed-Memory Transformer

III. Scan2CapMMT

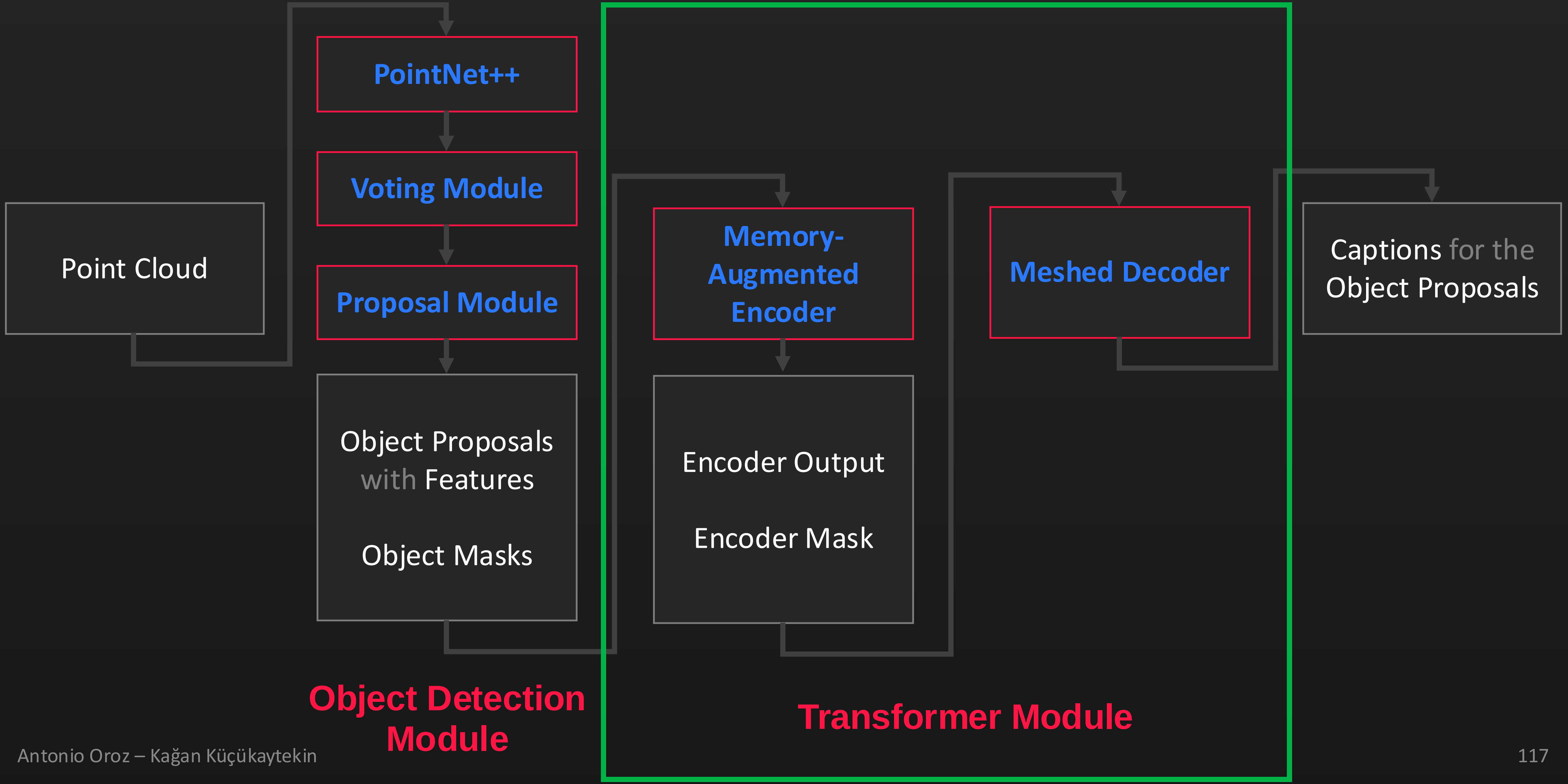
IV. Insights & First Results

V. Next Steps

III. Scan2CapMMT: Initial Architecture

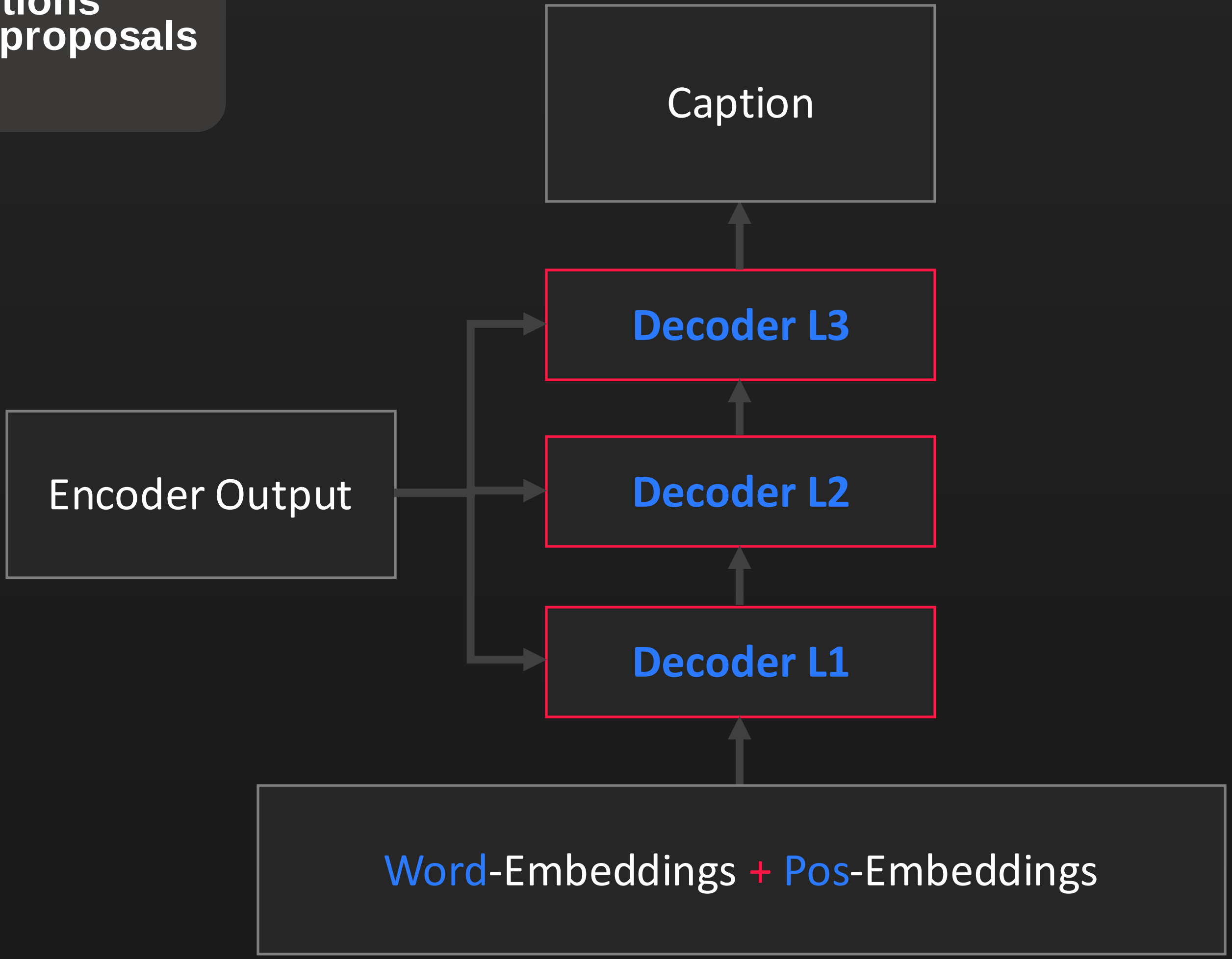


III. Scan2CapMMT: With MMT



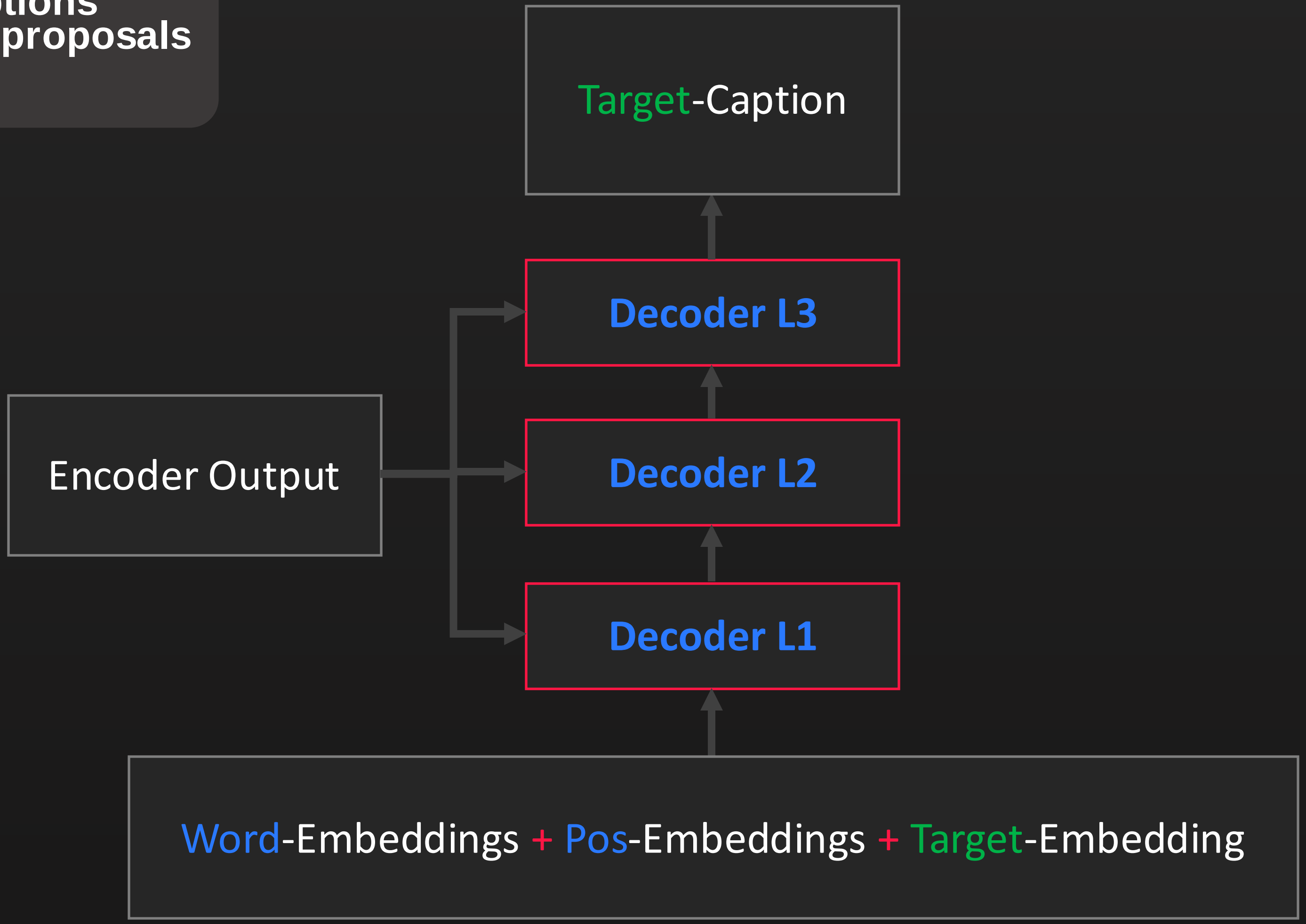
III. Scan2CapMMT: Challenges

Decoding Captions for multiple object proposals



III. Scan2CapMMT: Challenges

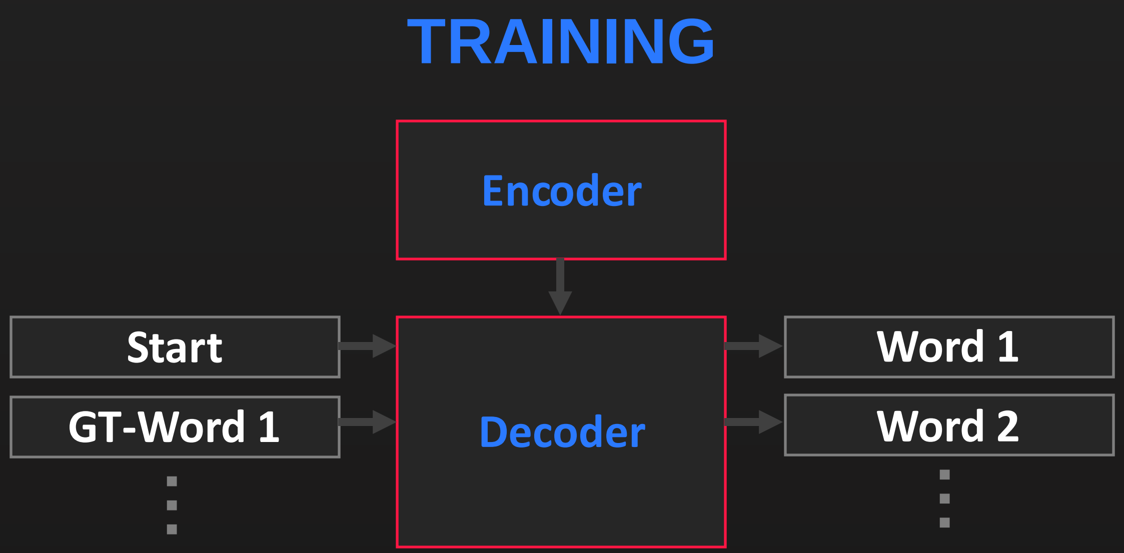
Decoding Captions for multiple object proposals

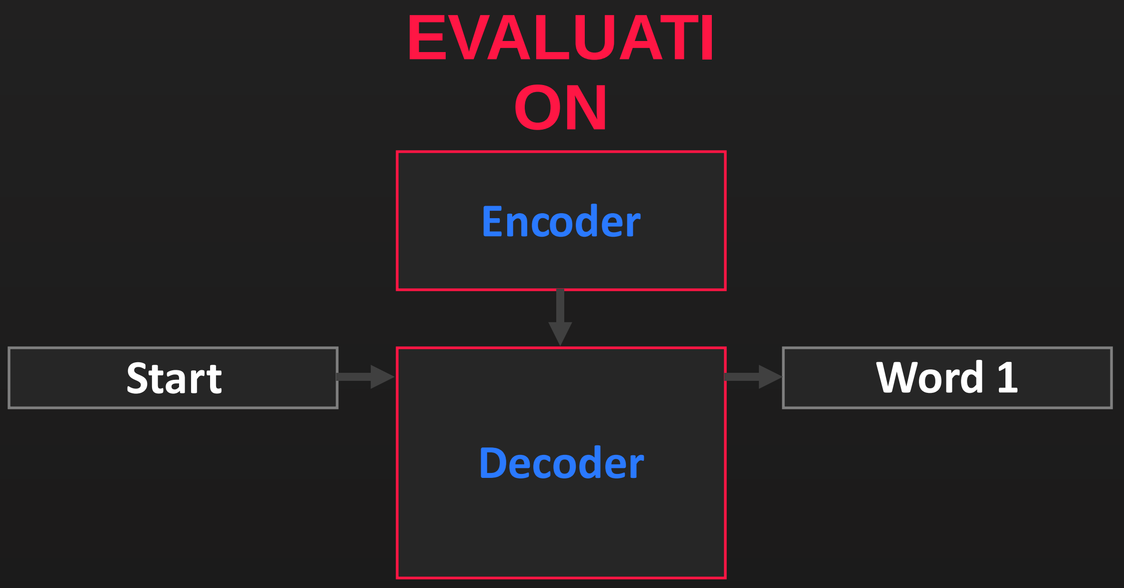


III. Scan2CapMMT: Challenges

Decoding Captions
for multiple object proposals

Caption-Generation
in Training and Evaluation

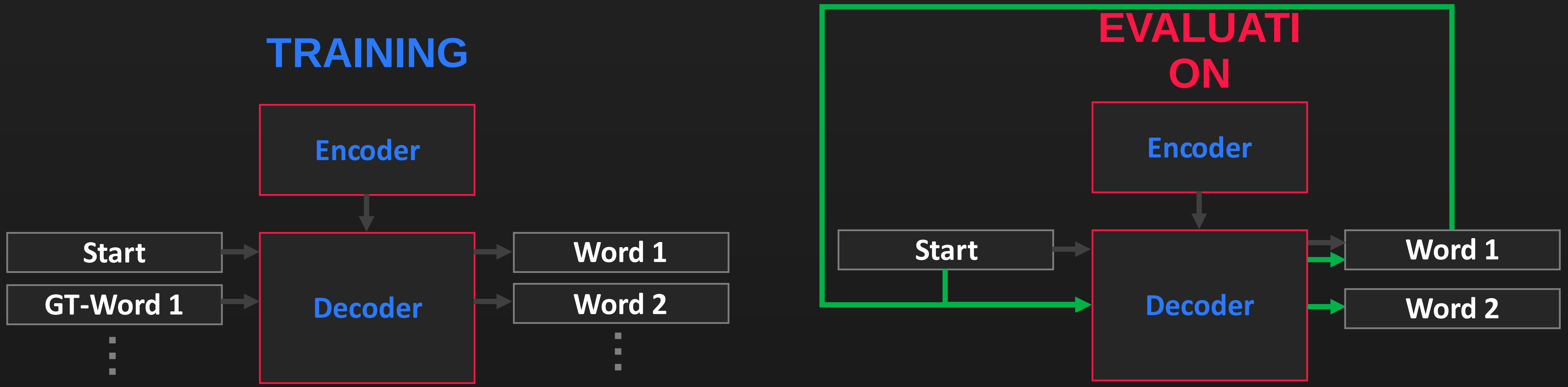




III. Scan2CapMMT: Challenges

Decoding Captions
for multiple object proposals

Caption-Generation
in Training and Evaluation



Scan2CapMMT

I. Scan2Cap

II. Meshed-Memory Transformer

III. Scan2CapMMT

IV. Insights & First Results

V. Next Steps

Scan2CapMMT

- I. Scan2Cap
- II. Meshed-Memory Transformer
- III. Scan2CapMMT
- IV. Insights & First Results
- V. Next Steps

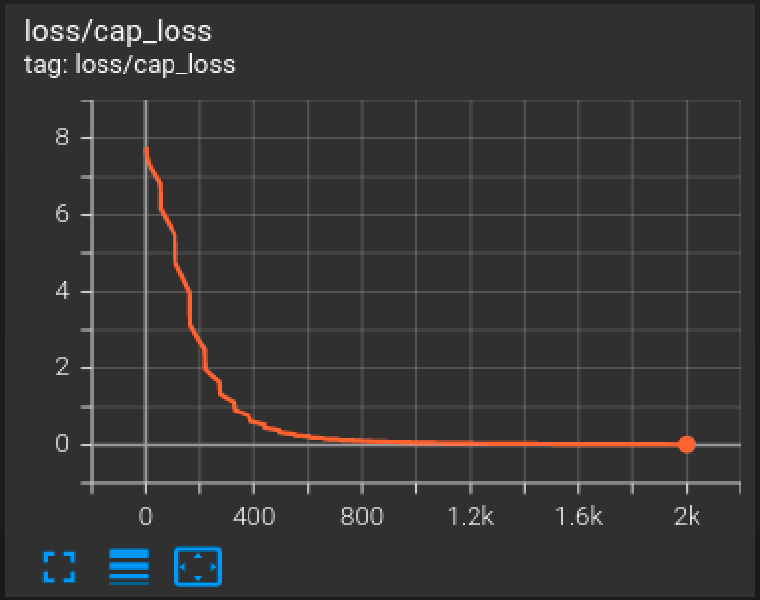
IV. Insights & First Results:

- **Parameters:** Scan2Cap MMT **7,830,308** vs **6,175,612** Scan2Cap
- **Dropout:** 0
- **Weight Decay:** 0
- **Learning Rate:** Changed from 1e-3 to 1e-4

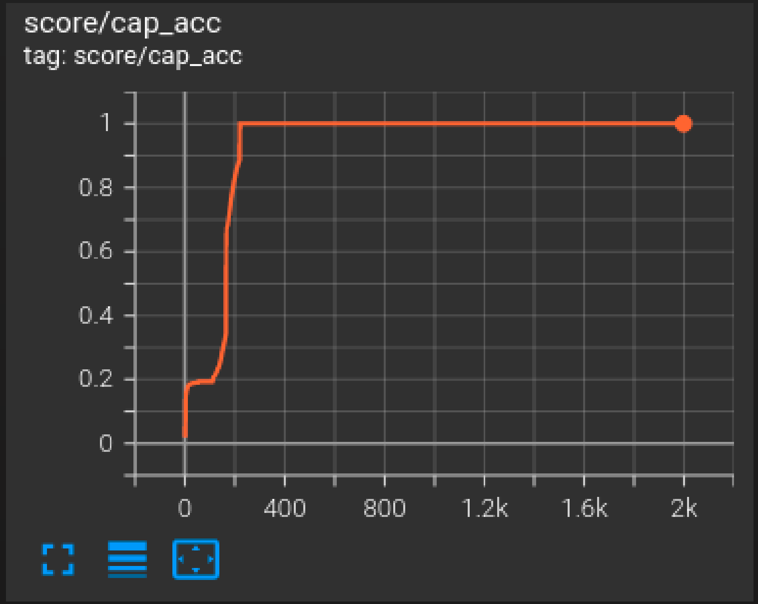
IV. Insights & First Results: Overfitting Results

1 SAMPLE 1 SCENE

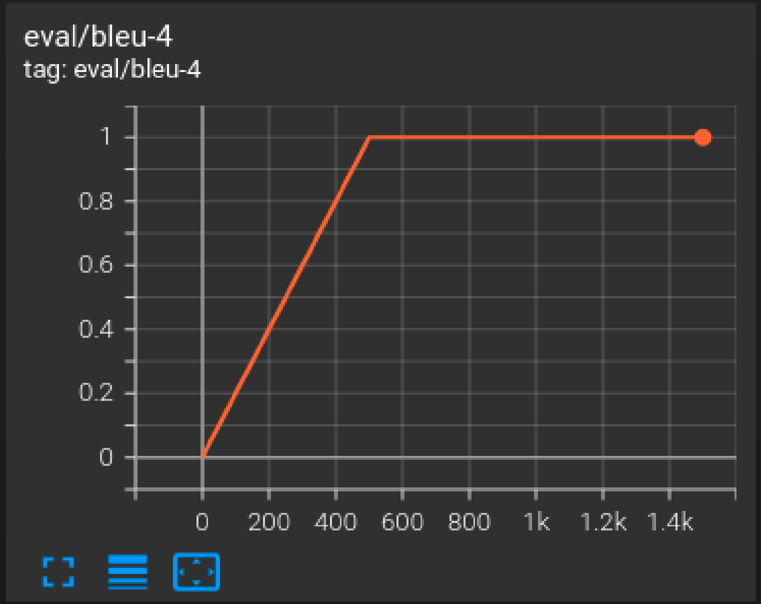
Caption Loss



Caption Accuracy



BLEU-4 Score

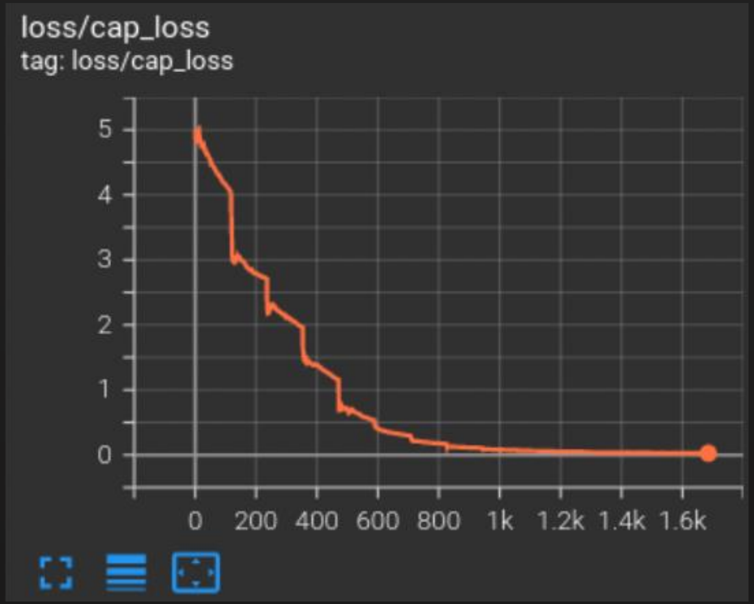


IV. Insights & First Results: Overfitting Results

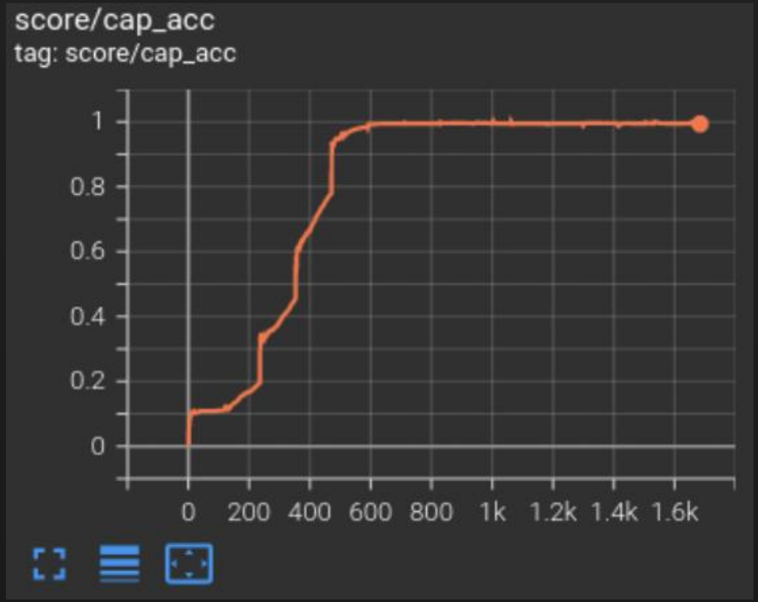
1 SAMPLE 1 SCENE

N SAMPLES 1 SCENE

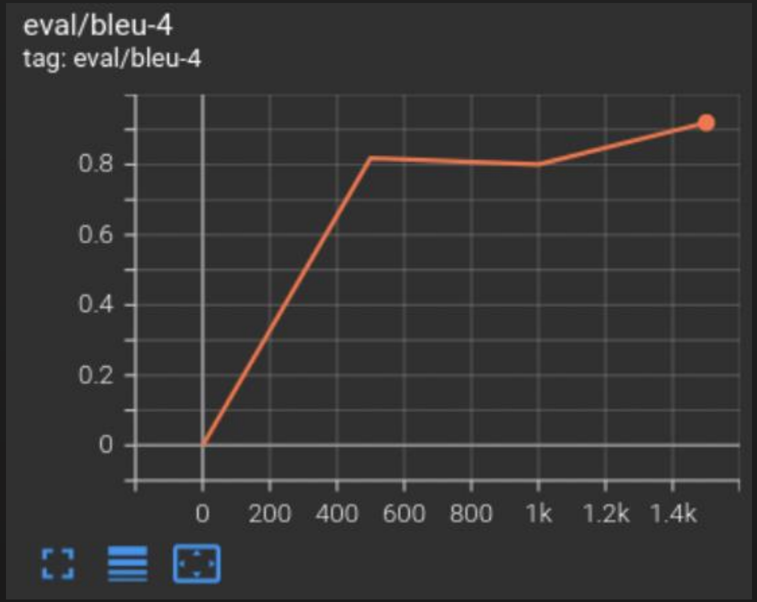
Caption Loss



Caption Accuracy



BLEU-4 Score



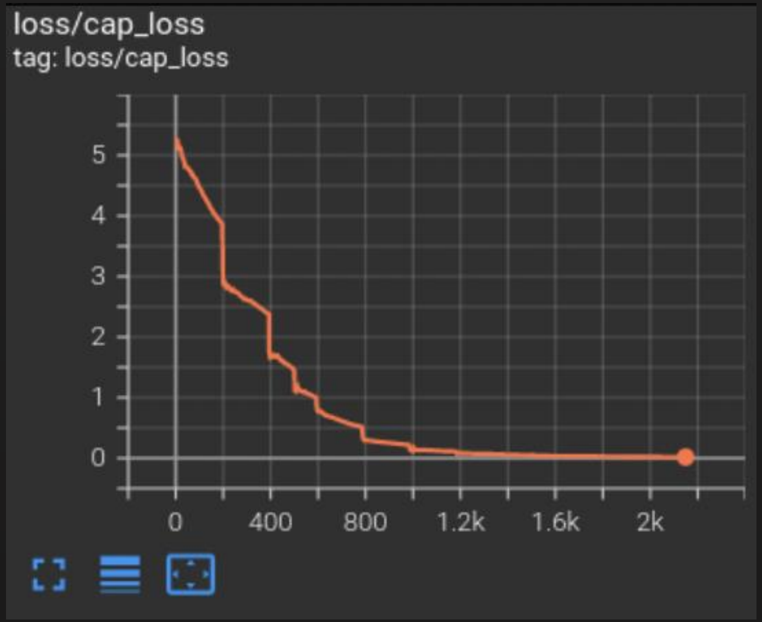
IV. Insights & First Results: Overfitting Results

1 SAMPLE 1 SCENE

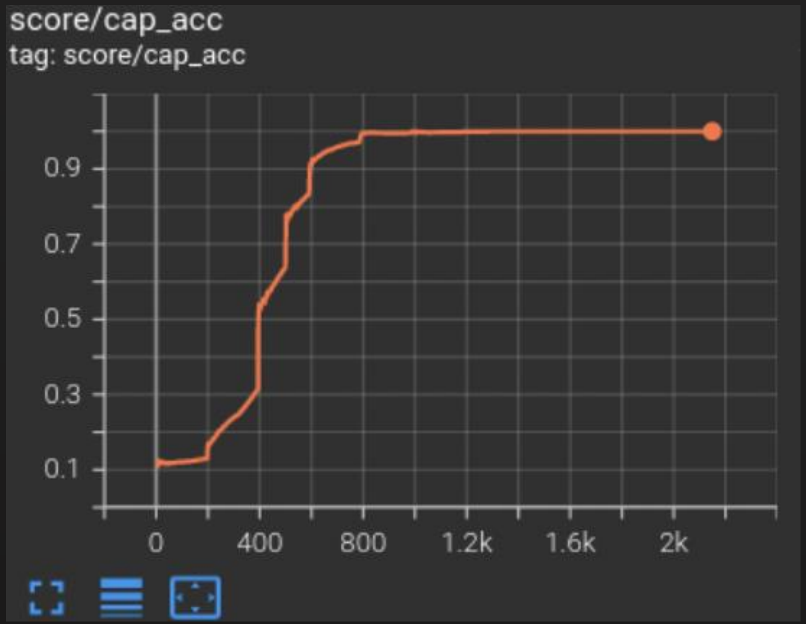
N SAMPLES 1 SCENE

N SAMPLES M SCENES

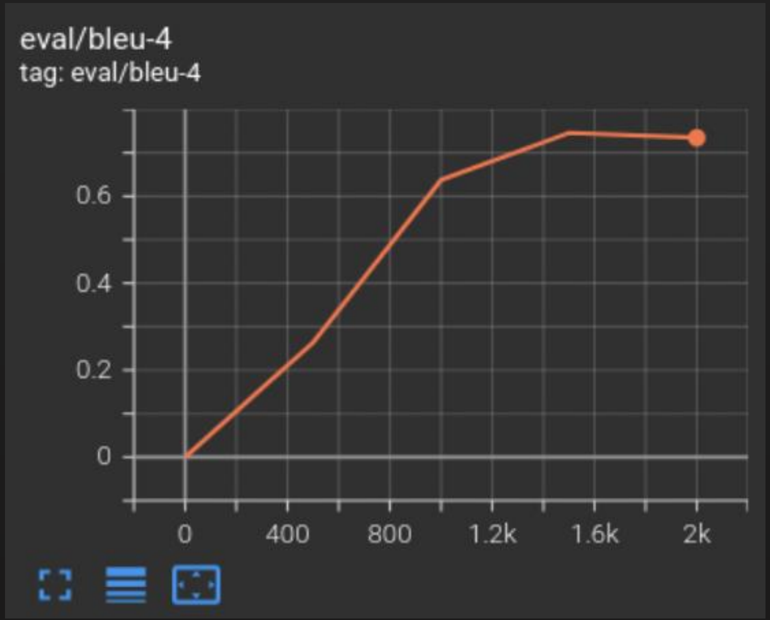
Caption Loss



Caption Accuracy



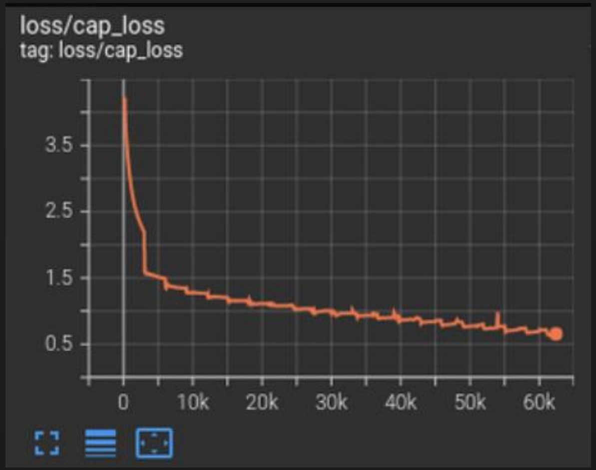
BLEU-4 Score



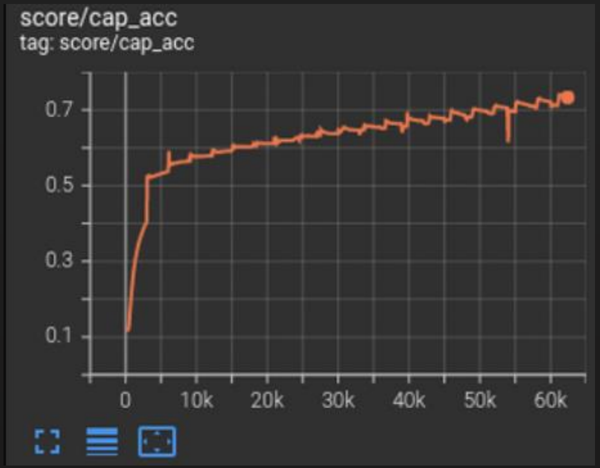
IV. Insights & First Results: Training on the whole Dataset

Losses & Accuracies

Caption Loss



Caption Accuracy

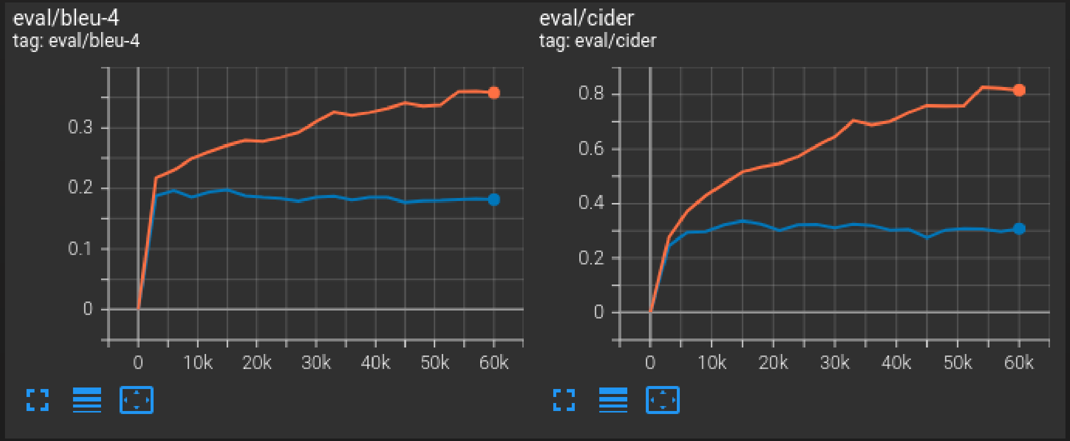


IV. Insights & First Results: Training on the whole Dataset

Losses & Accuracies

Evaluation

Scan2CapMMT



Scan2Cap

Scan2CapMMT

I. Scan2Cap

II. Meshed-Memory Transformer

III. Scan2CapMMT

IV. Insights & First Results

V. Next Steps

Scan2CapMMT

V. Next Steps

BEAM SEARCH

Instead of generating one sentence for an object proposal, generate multiple sentences in parallel and choose the final sentence with log propobabilities.

V. Next Steps

BEAM SEARCH

REINFOCEMENT LEARNING

After pretraining on the Cross-Entropy loss,
use Reinforcement Learning with CIDEr-D as a reward
to train the model.

V. Next Steps

BEAM SEARCH | REINFOCEMENT LEARNING | **HYPERPARAMETER TUNING**

- Internal Dimensions of MMT
- Number of Proposals
- # Decoder-/Encoder-Layers
- Schedules
- Learning Rate
- Weight Decay

...

V. Next Steps

BEAM SEARCH | REINFOCEMENT LEARNING | HYPERPARAMETER TUNING | **GROUP-FREE TRANSFORMER**

Replace the current detection module
with the Group-Free 3D Object Detection via Transformers module
proposed by Liu et al.

V. Next Steps

BEAM SEARCH | REINFOCEMENT LEARNING | HYPERPARAMETER TUNING | GROUP-FREE TRANSFORMER | **AoA**

MMT currently uses Dot-Product Attention
which we could replace with Attention on Attention

Scan2CapMMT

I. Scan2Cap

II. Meshed-Memory Transformer

III. Scan2Cap with MMT

IV. Insights & First Results

V. Next Steps

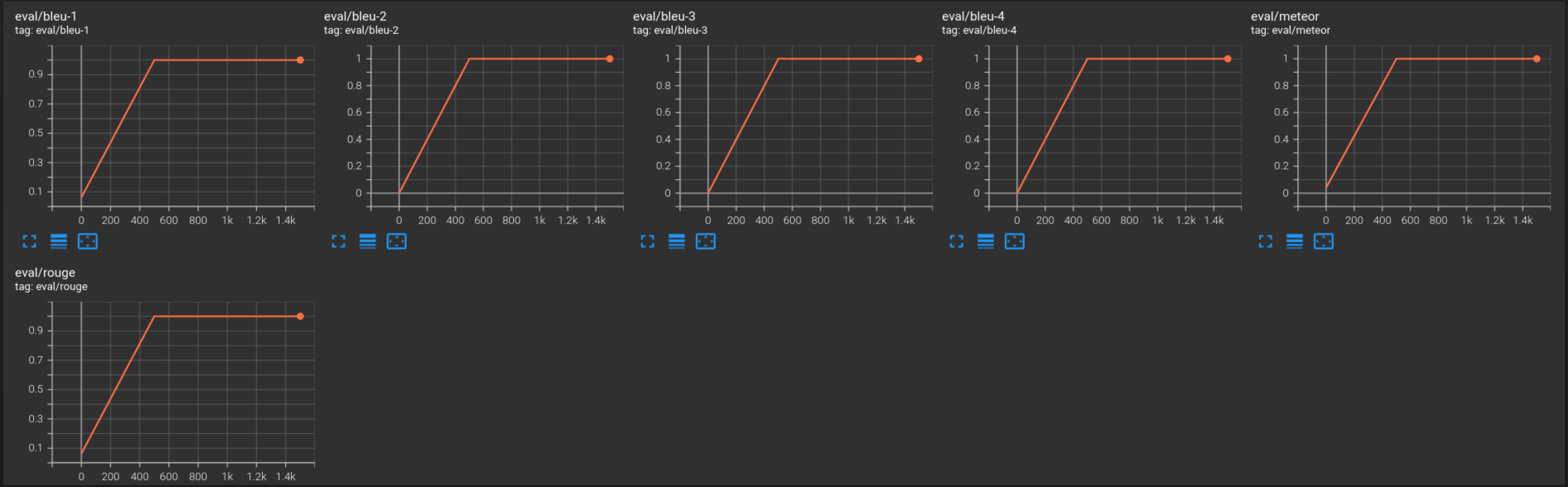
Scan2CapMMT

THANK YOU FOR YOUR ATTENTION :D

BACKUP: Overfitting Results

1 SAMPLE 1 SCENE

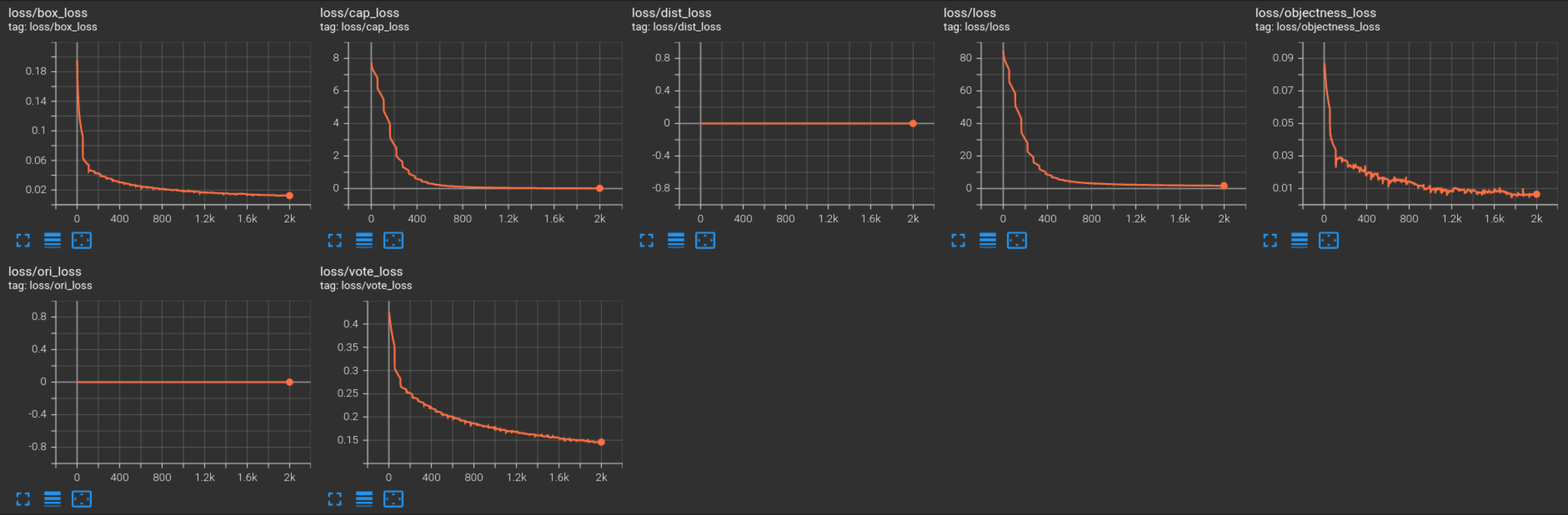
Evaluation



BACKUP: Overfitting Results

1 SAMPLE 1 SCENE

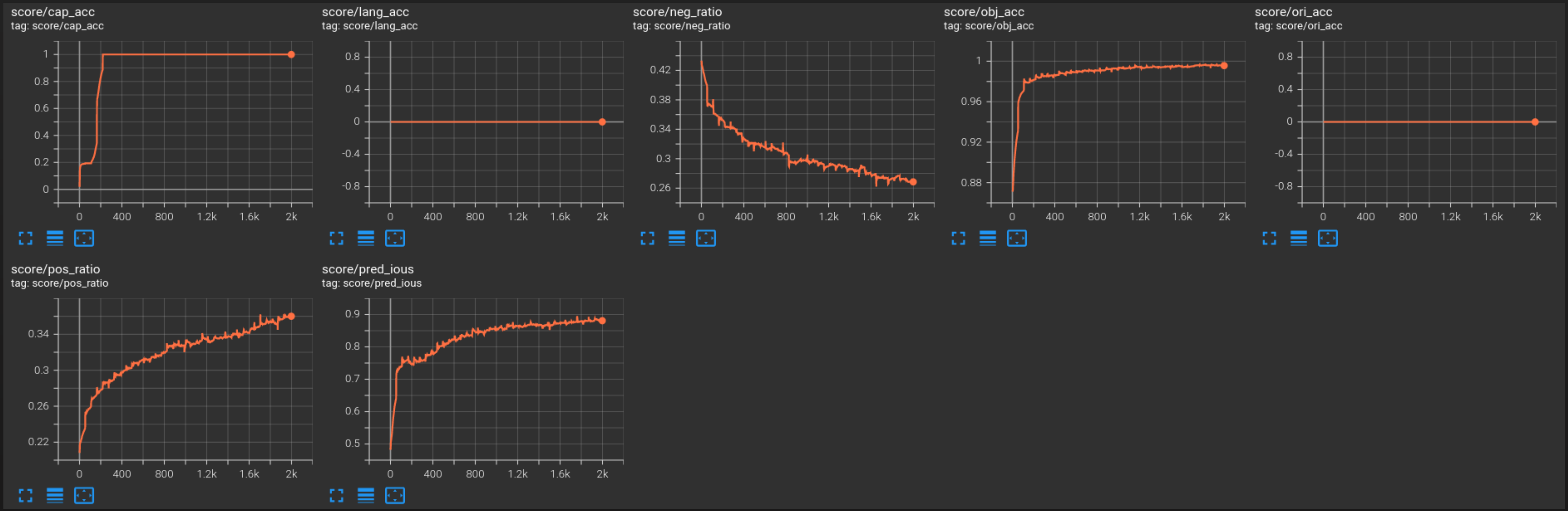
Losses



BACKUP: Overfitting Results

1 SAMPLE 1 SCENE

Accuracies

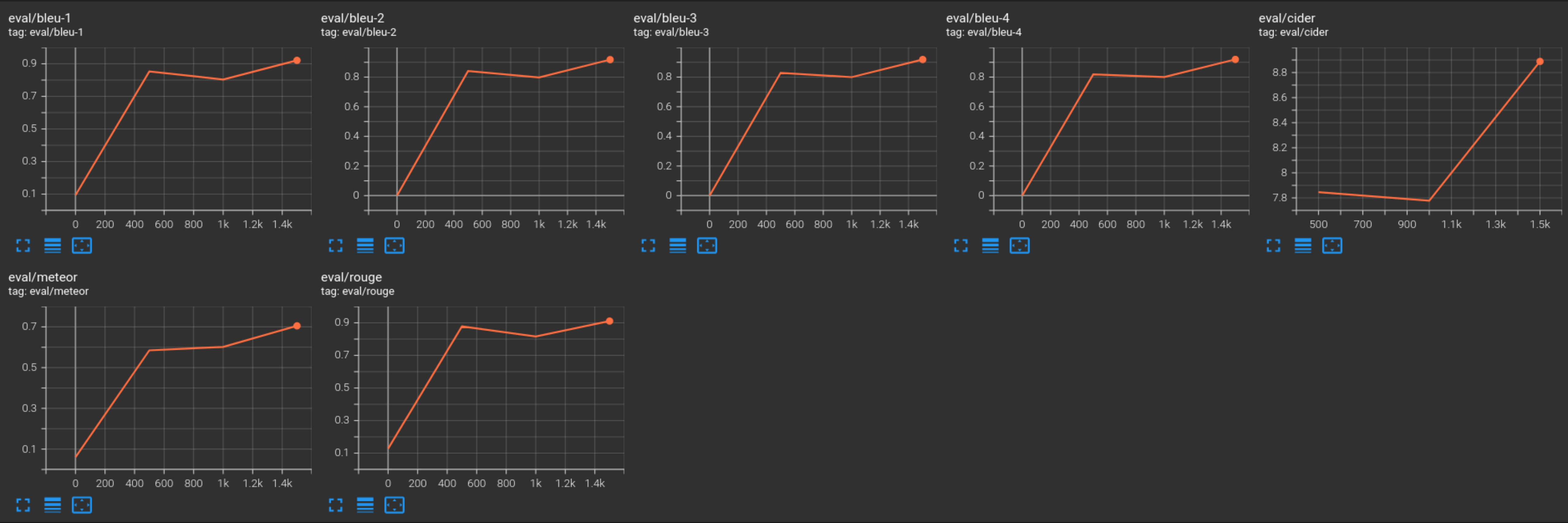


BACKUP: Overfitting Results

1 SAMPLE 1 SCENE

N SAMPLES 1 SCENE

Evaluation

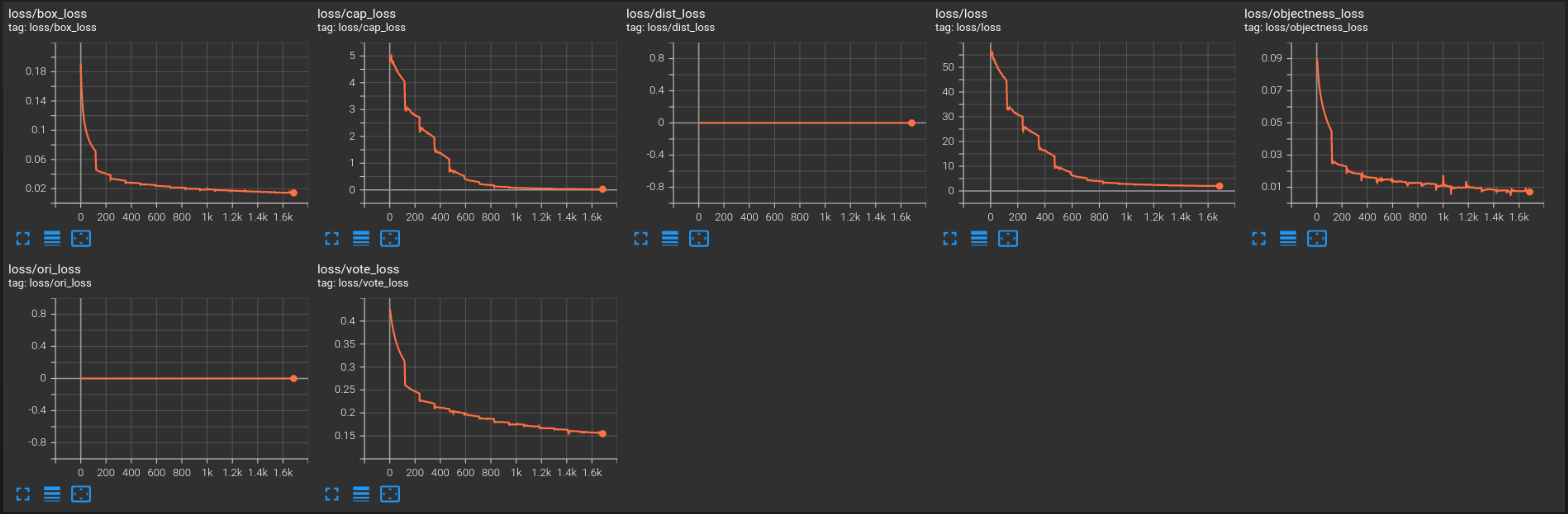


BACKUP: Overfitting Results

1 SAMPLE 1 SCENE

N SAMPLES 1 SCENE

Losses

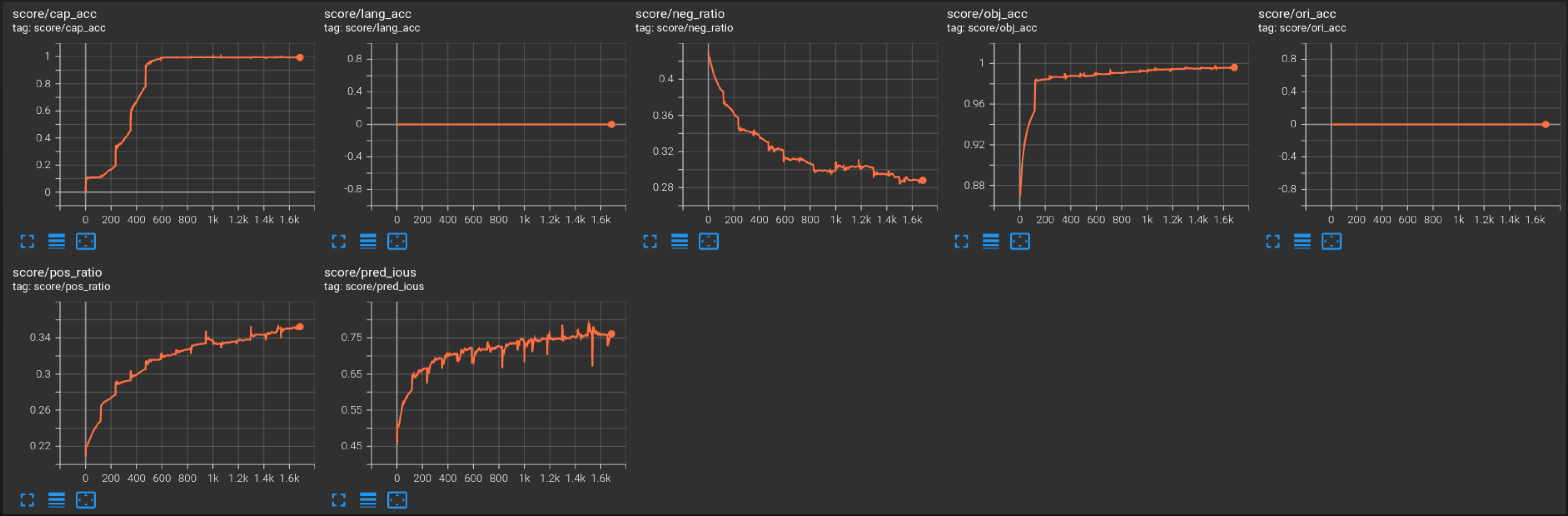


BACKUP: Overfitting Results

1 SAMPLE 1 SCENE

N SAMPLES 1 SCENE

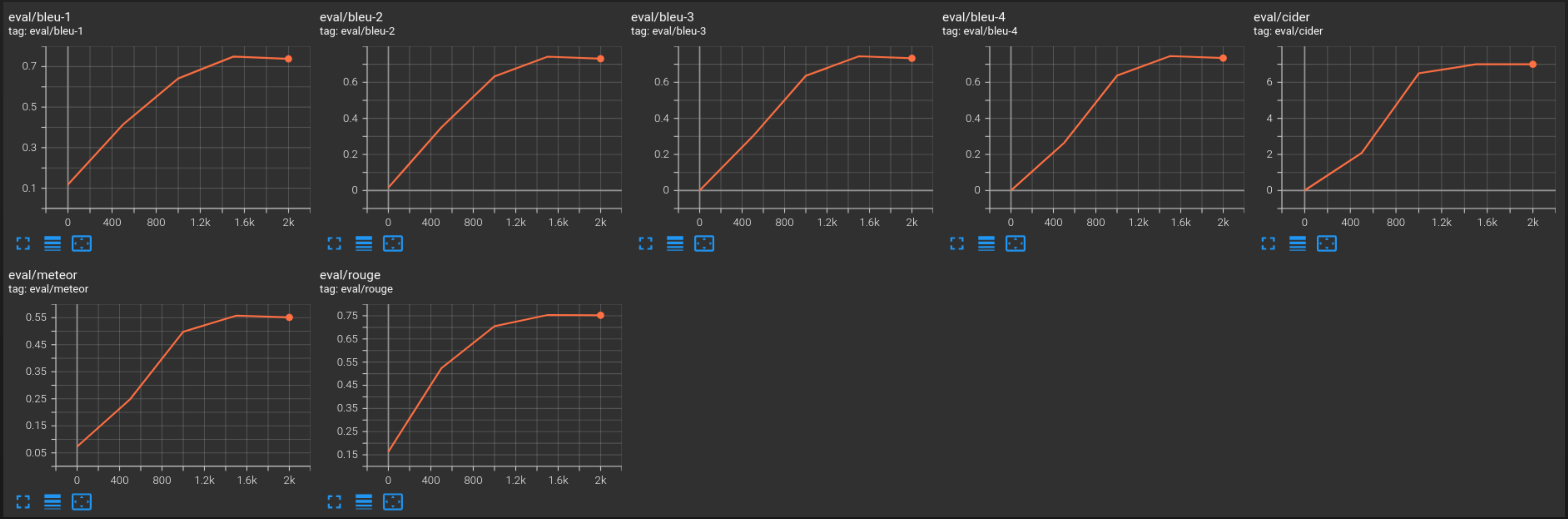
Accuracies



BACKUP: Overfitting Results

1 SAMPLE 1 SCENE | N SAMPLES 1 SCENE | **N SAMPLES M SCENES**

Evaluation



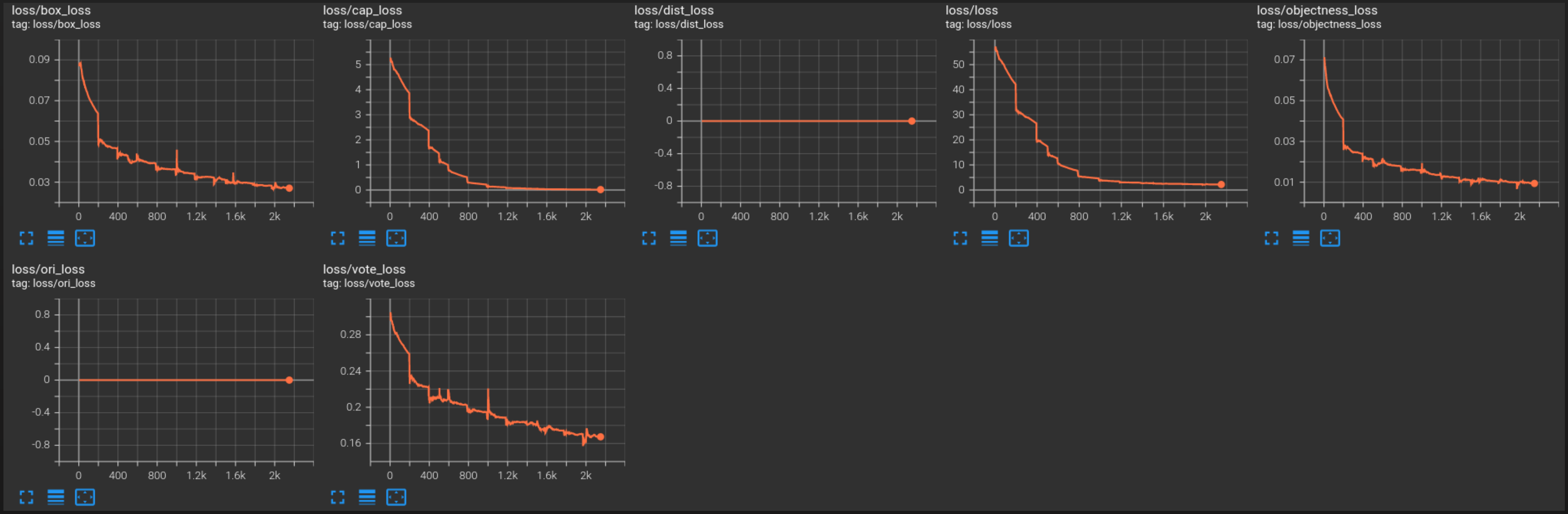
BACKUP: Overfitting Results

1 SAMPLE 1 SCENE

N SAMPLES 1 SCENE

N SAMPLES M SCENES

Losses



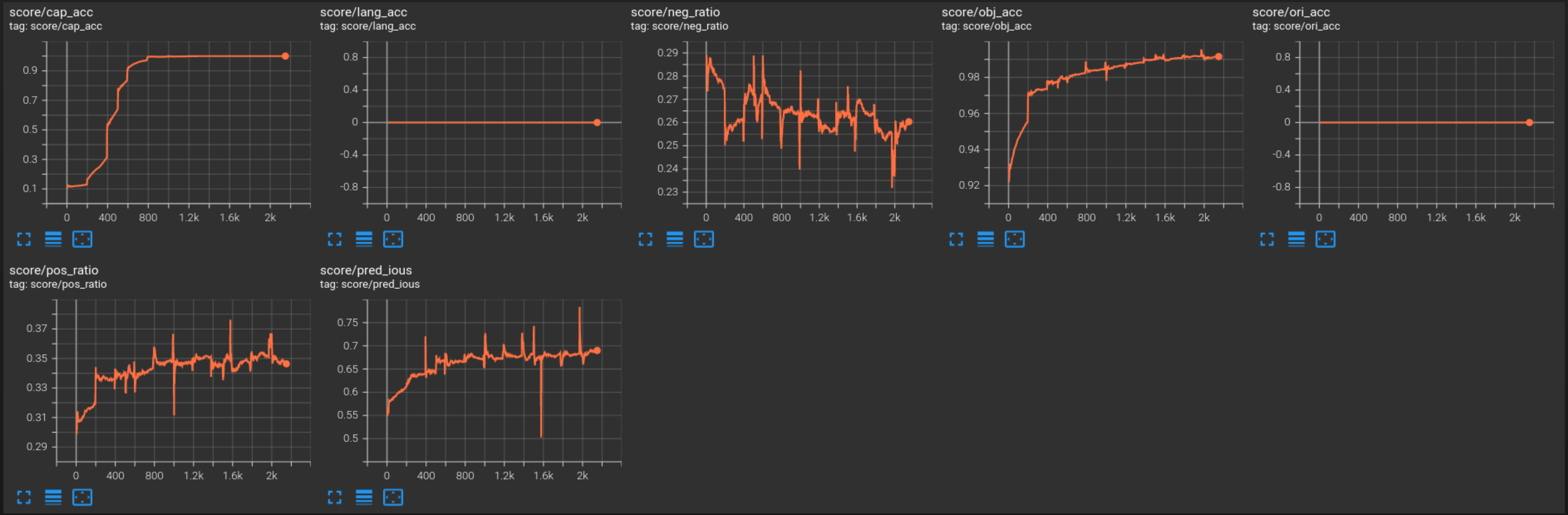
BACKUP: Overfitting Results

1 SAMPLE 1 SCENE

N SAMPLES 1 SCENE

N SAMPLES M SCENES

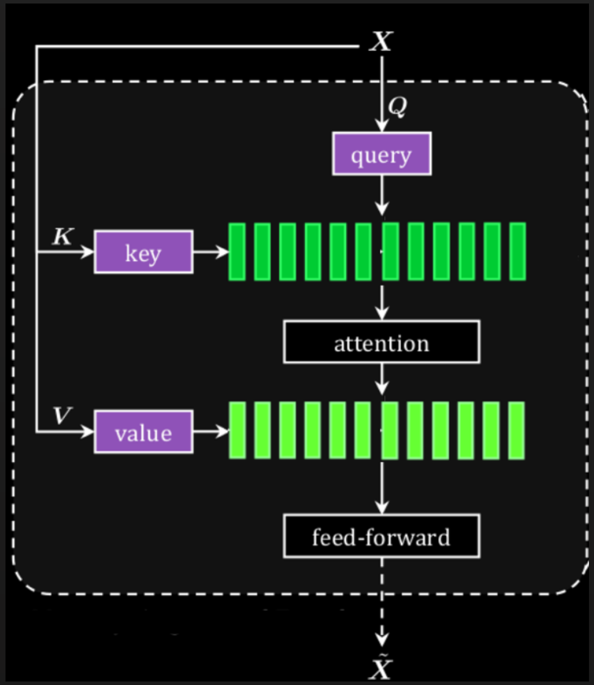
Accuracies



II. Meshed-Memory Transformer

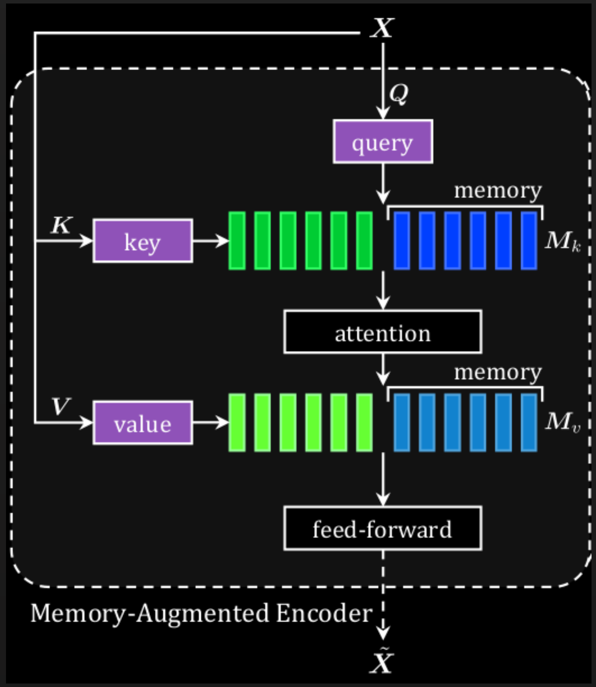
Encoder Layer 1

Attention



Develop and maintaion a-priori knowledge in persistent memory vectors

Memory Augmented Attention

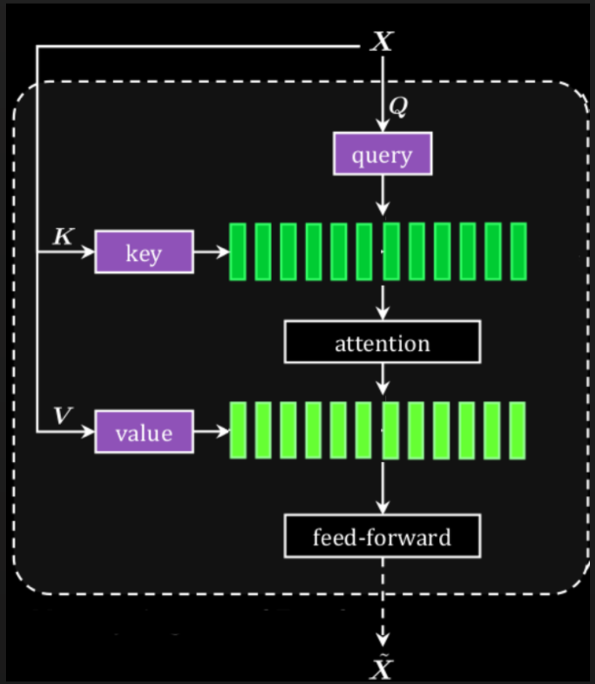


II. Meshed-Memory Transformer: Attention

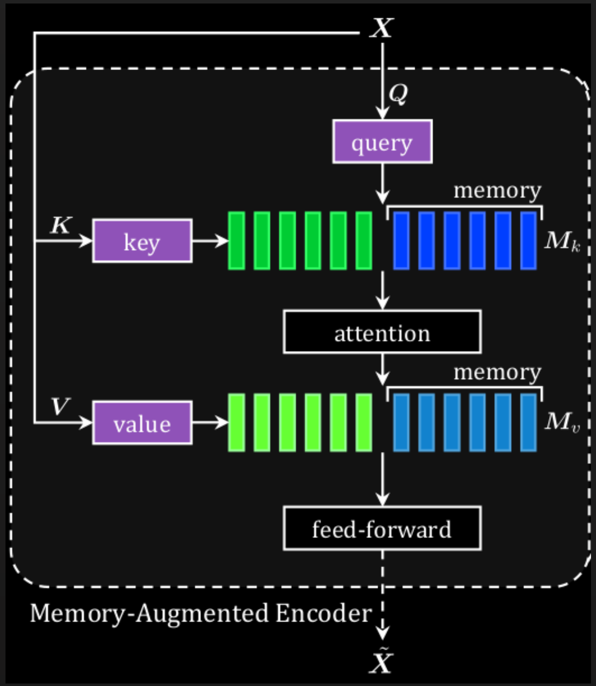
- Fully attentive.
- Scaled dot-product attention, without recurrence.
- Self attention in decoders
- Cross-attention bMeshedetween decoder and encoder
- Masked self-attention between decoders

II. Meshed-Memory Transformer: Encoder

Attention → Memory Augmented Attention



Develop and maintaion a-priori knowledge in persistent memory vectors



II. Meshed-Memory Transformer: Decoder