

CIC-BART-SSA: Controllable Image Captioning with Structured Semantic Augmentation

Kalliopi Basioti, Mohamed A. Abdelsalam, Federico Fancellu, Vladimir Pavlovic, Afsaneh Fazly



EUROPEAN CONFERENCE ON COMPUTER VISION

M I L A N O
2 0 2 4

Controllable Image Captioning (CIC)

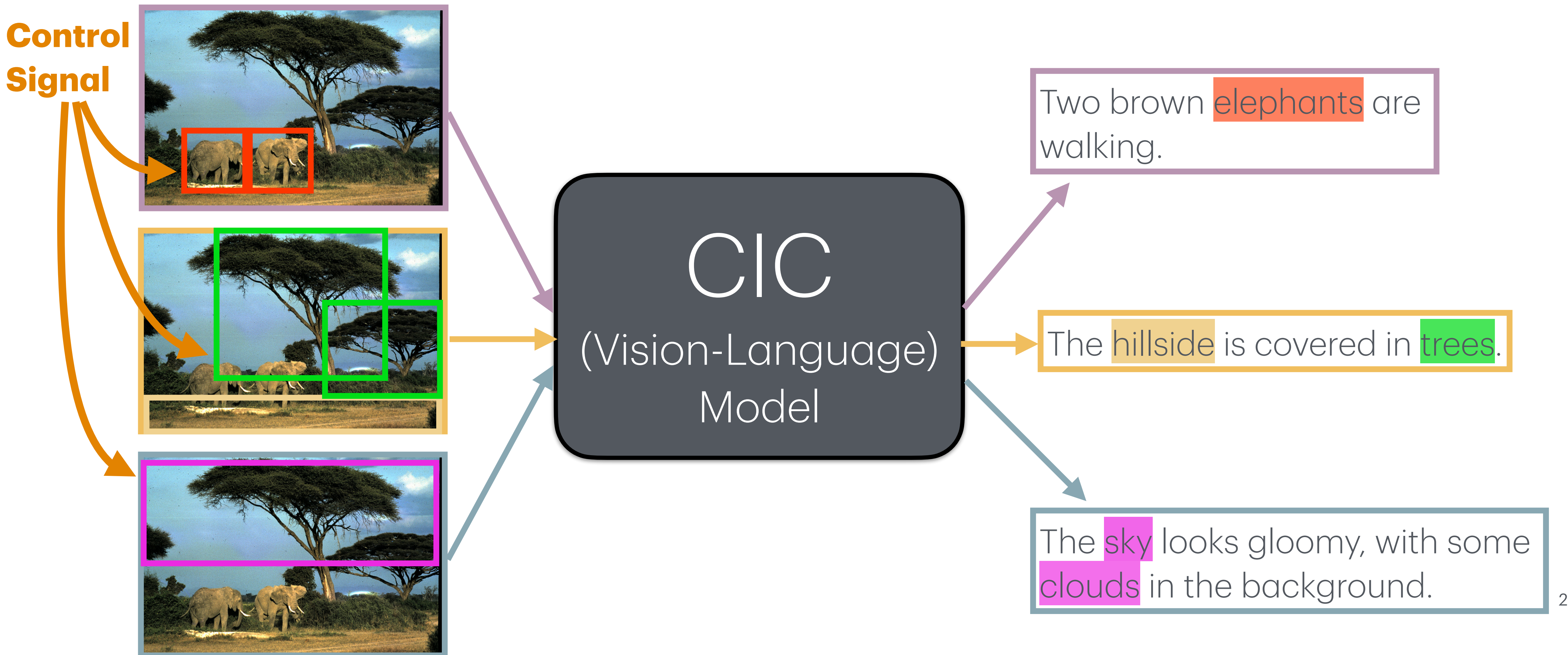
The CIC task.

**Control
Signal**



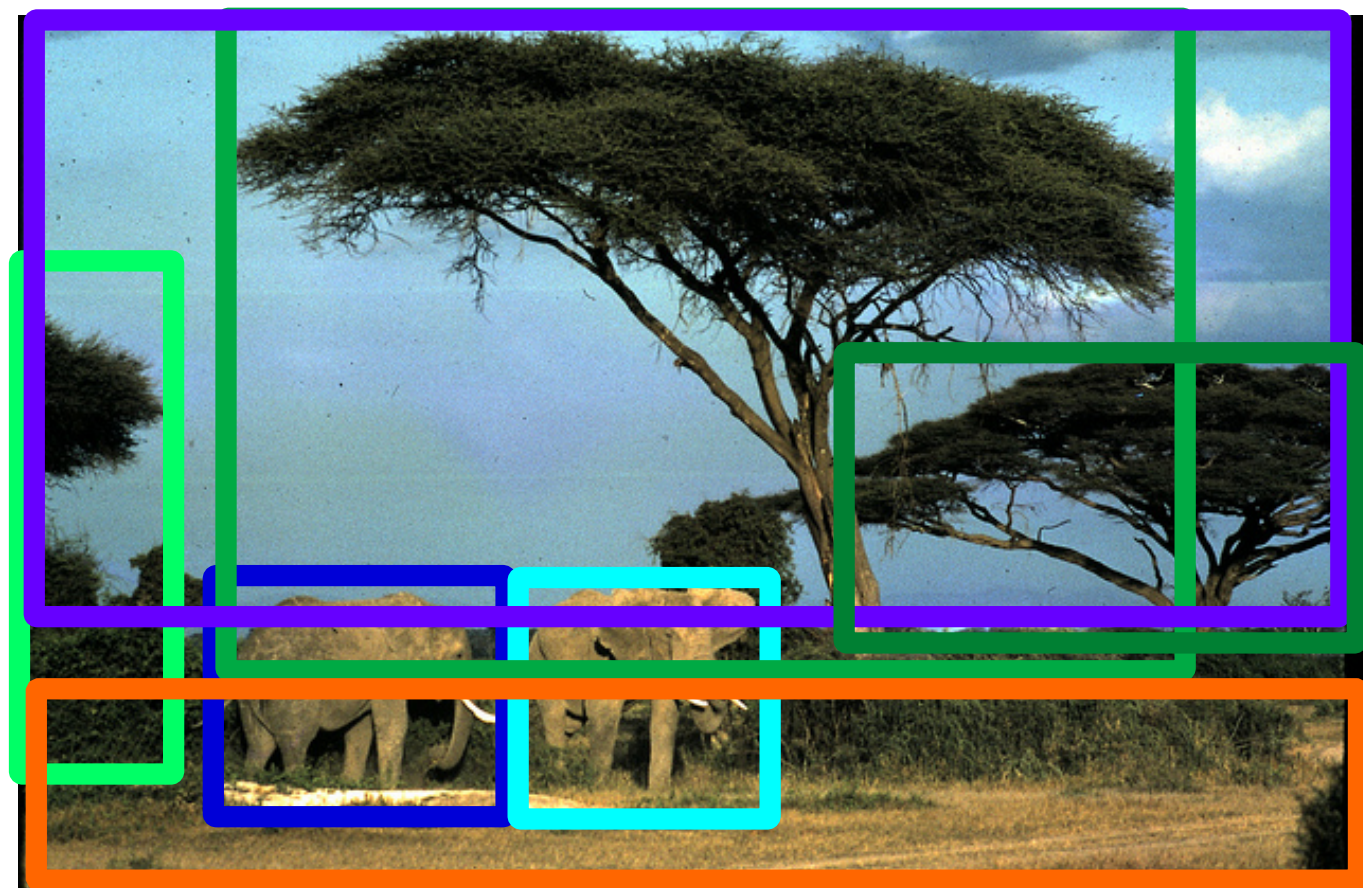
Controllable Image Captioning (CIC)

The CIC task.



Controllable Image Captioning (CIC)

Do existing datasets help CIC reach its goals?



Original Captions

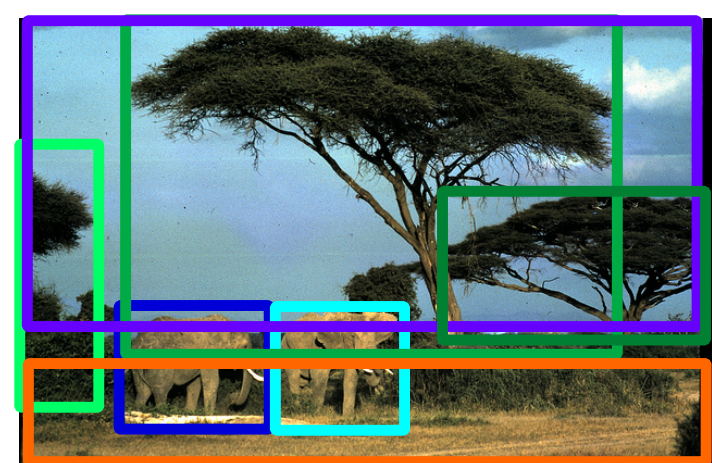
A couple of **elephants** standing next to **trees**.
A couple of **elephants** standing by some **trees**.
A couple **elephants** walking by a **tree** after **sunset**.
Two **elephants** are standing in the **grass** near a **tree**.
Two **elephants** are standing by the **trees** in the wild.

Samples Generated from CIC-BART-SSA

There is **grass** near a **tree**.
There is a **tree** near the **grass**.
The **hillside** is covered in **trees**.
A **field** next to a big **tree** is open.
The **grass** under the **trees** is lush green.
A large **elephant** standing in a field near a **tree**.
Two **elephants** walking under a **tree** in the **sunset**.
Two **elephants** standing in the **grass** next to a **tree**.
Two **elephants** standing next to a **tree** and a blue **sky**.
A big **elephant** standing next to a **tree** in a **field** of grass.
An **elephant** standing next to a **tree** with a blue **sky** behind it.
Two **elephants** standing in the **grass** near a **tree** and a blue **sky**.
Two **elephants** standing next to each other on a lush green **hillside** next to a **forest**.
Two **elephants** standing next to a **tree** in a **field** with a blue **sky** and **trees** behind them.
Two **elephants** standing next to each other on a grass covered **field** next to a lush green **forest**.

Controllable Image Captioning (CIC)

Do existing datasets help CIC reach its goals?

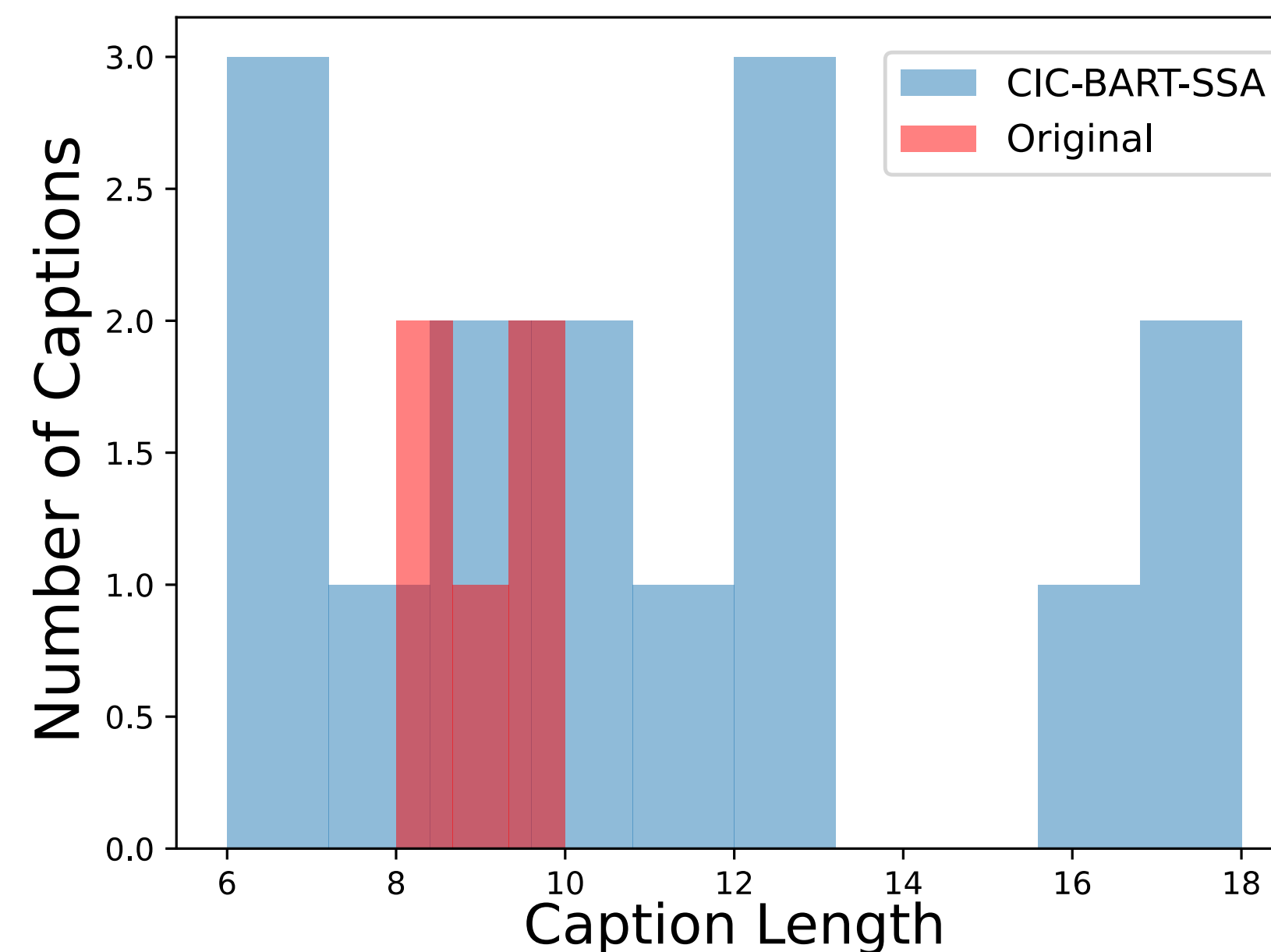
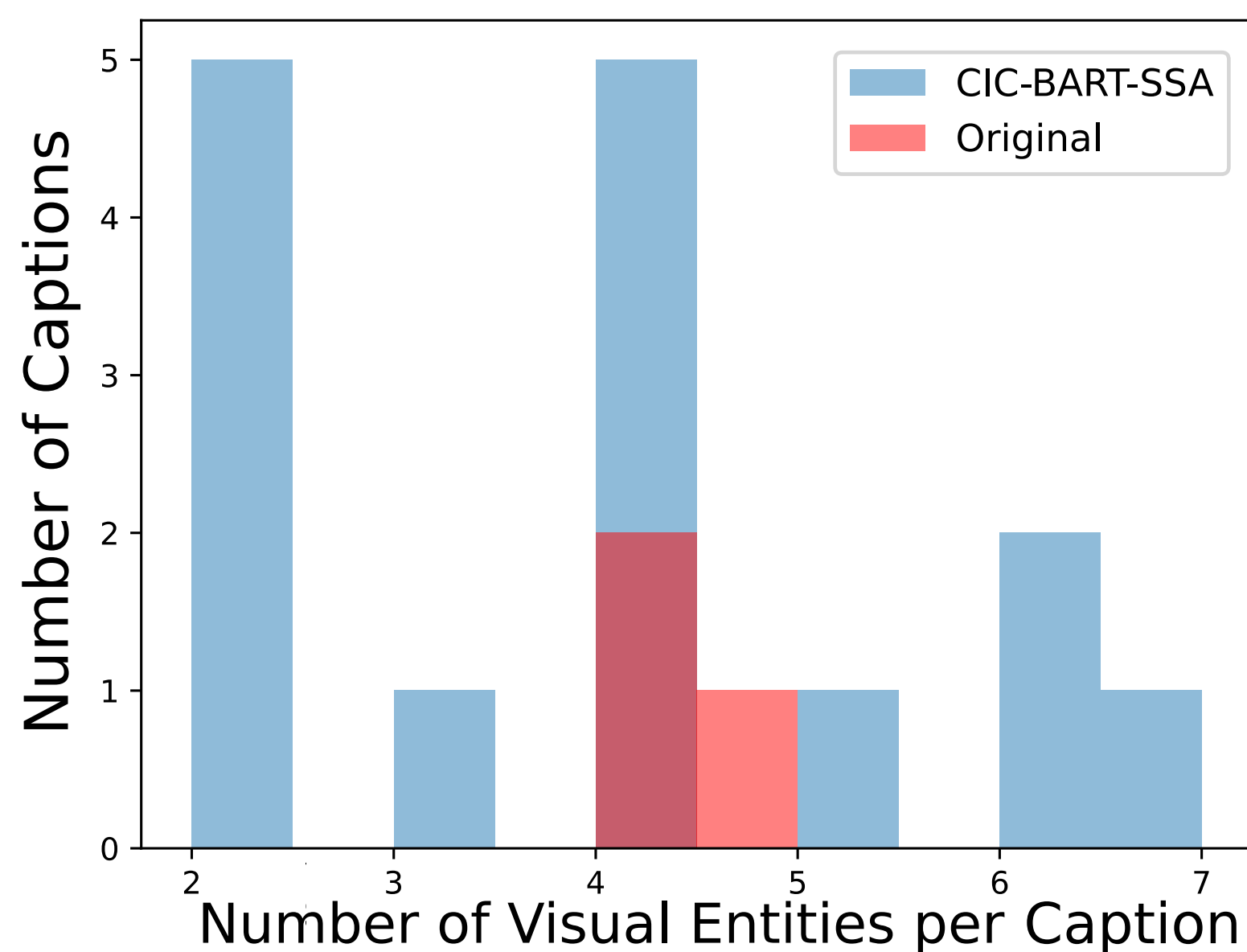


Original Captions

- A couple of elephants standing next to trees.
- A couple of elephants standing by some trees.
- A couple elephants walking by a tree after sunset.
- Two elephants are standing in the grass near a tree.
- Two elephants are standing by the trees in the wild.

Samples Generated from CIC-BART-SSA

- There is grass near a tree.
- There is a tree near the grass.
- The hillside is covered in trees.
- A field next to a big tree is open.
- The grass under the trees is lush green.
- A large elephant standing in a field near a tree.
- Two elephants walking under a tree in the sunset.
- Two elephants standing in the grass next to a tree.
- Two elephants standing next to a tree and a blue sky.
- A big elephant standing next to a tree in a field of grass.
- An elephant standing next to a tree with a blue sky behind it.
- Two elephants standing in the grass near a tree and a blue sky.
- Two elephants standing next to each other on a lush green hillside next to a forest.
- Two elephants standing next to a tree in a field with a blue sky and trees behind them.
- Two elephants standing next to each other on a grass covered field next to a lush green forest.



Controllable Image Captioning (CIC)

Our Goals.

Controllable Image Captioning (CIC)

Our Goals.

- **Regarding CIC Datasets**

- Goal — **Dataset Diversity**
 - Spatially diverse Control Signals
 - Linguistically diverse Captions

Controllable Image Captioning (CIC)

Our Goals.

• Regarding CIC Datasets

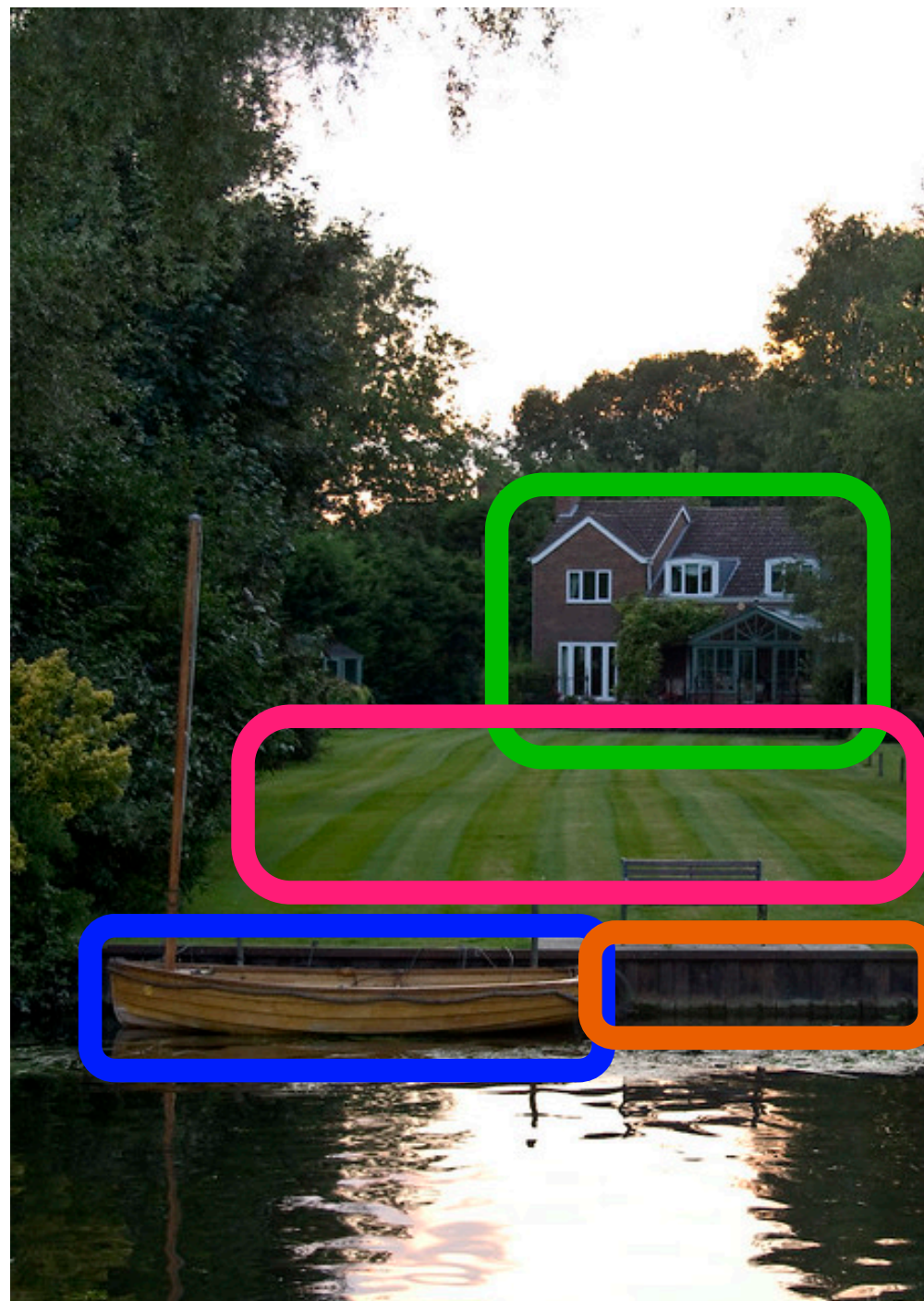
- Goal — **Dataset Diversity**
 - Spatially diverse Control Signals
 - Linguistically diverse Captions

• Regarding CIC Models

- Goal — **Performance & Simplicity**
 - State-of-the-art performance where generated captions are
 - Faithful to control signal (controllability)
 - Linguistically Coherent
 - Linguistically Diverse
 - Simple, user-friendly control signals

Structured Semantic Augmentations (SSA)

A novel fully-automatic data augmentation approach suitable for CIC.

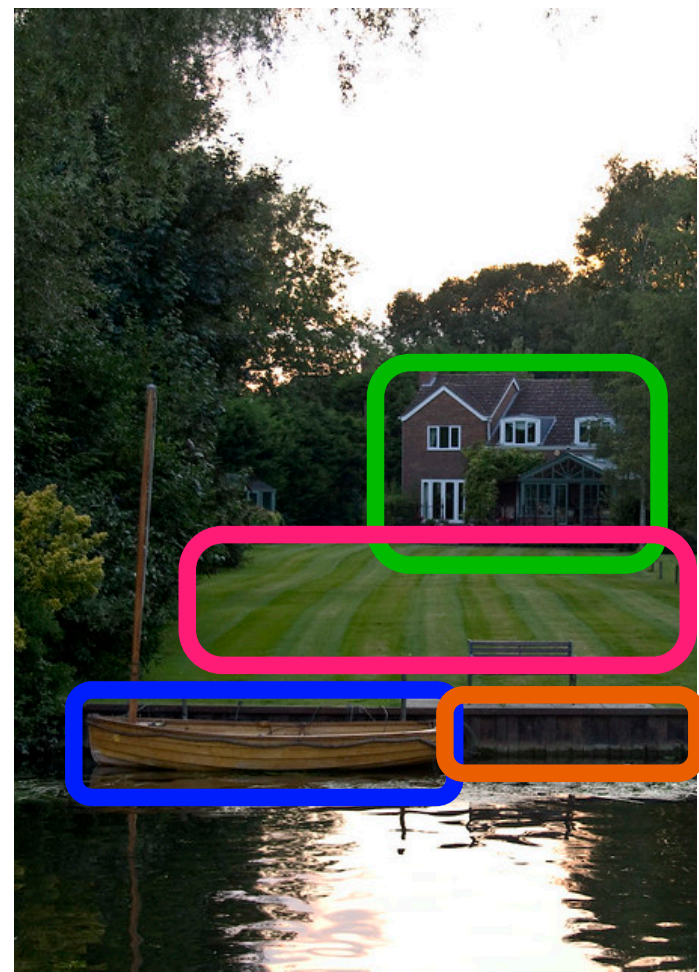


Original Captions

- (1) a **house** with a freshly mowed **lawn** is preceded by a small **dock** with a **boat**.
- (2) a **boat** sits in the water in front of a brick two story **house**.
- (3) a **boat** is docked in the water near a large **house**.
- (4) a view of a **house** from across the water.
- (5) a large body of water sitting in front of a **house** and green **lawn**.

Structured Semantic Augmentations (SSA)

A novel fully-automatic data augmentation approach suitable for CIC.

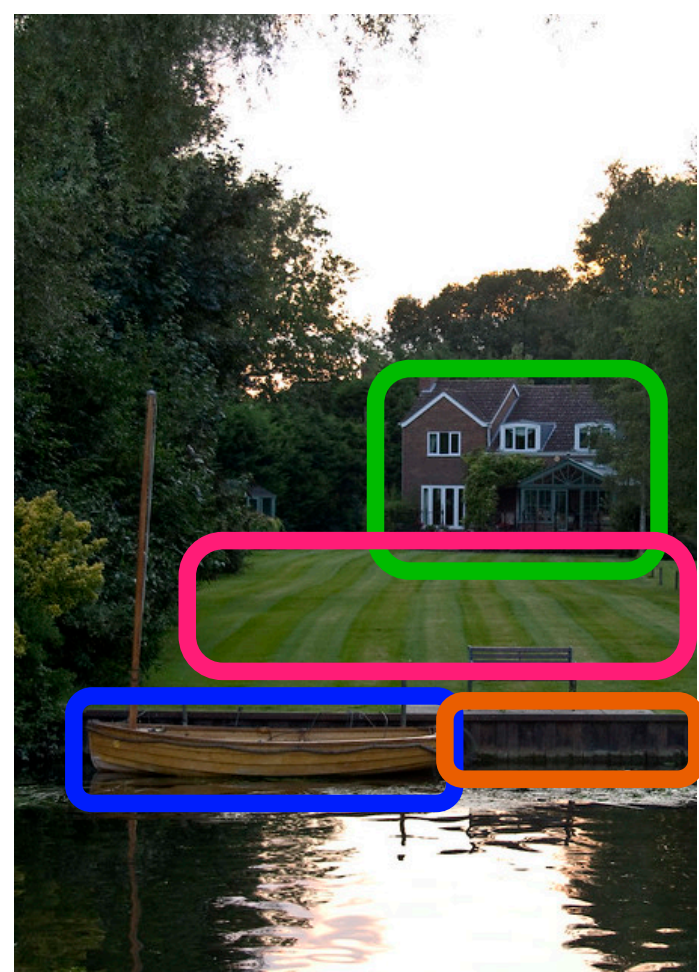


Original Captions

- (1) a **house** with a freshly mowed **lawn** is preceded by a small **dock** with a **boat**.
- (2) a **boat** sits in the water in front of a brick two story **house**.
- (3) a **boat** is docked in the water near a large **house**.
- (4) a view of a **house** from across the water.
- (5) a large body of water sitting in front of a **house** and green **lawn**.

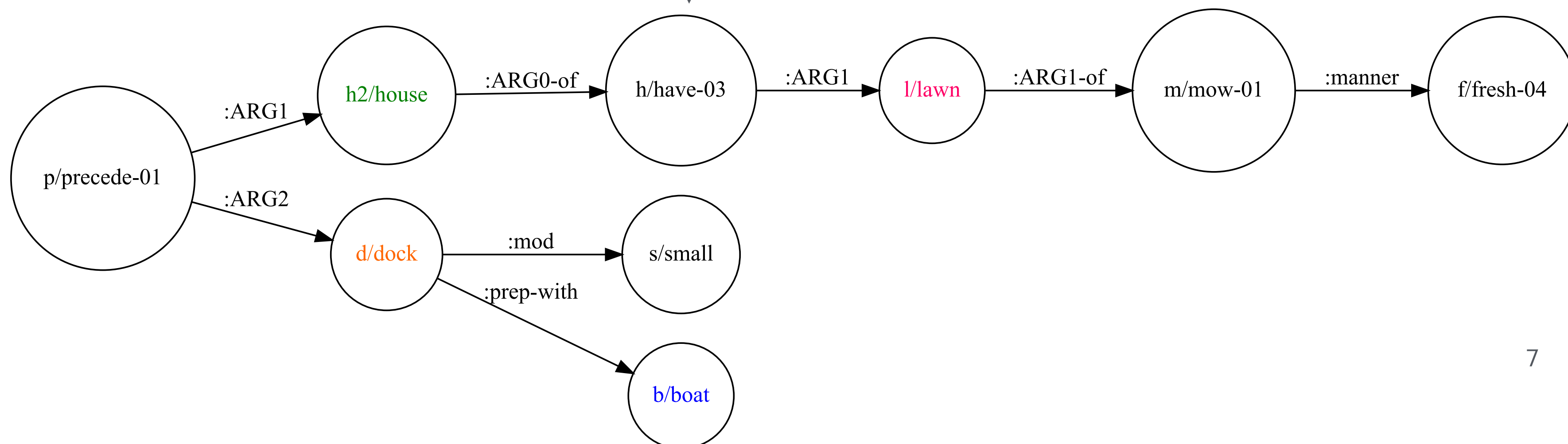
Structured Semantic Augmentations (SSA)

A novel fully-automatic data augmentation approach suitable for CIC.



(1) a **house** with a freshly mowed **lawn** is preceded by a small **dock** with a **boat**.

Text to AMR
+
AMR Visual Grounding



Original Captions

(1) a **house** with a freshly mowed **lawn** is preceded by a small **dock** with a **boat**.

(2) a **boat** sits in the water in front of a brick two story **house**.

(3) a **boat** is docked in the water near a large **house**.

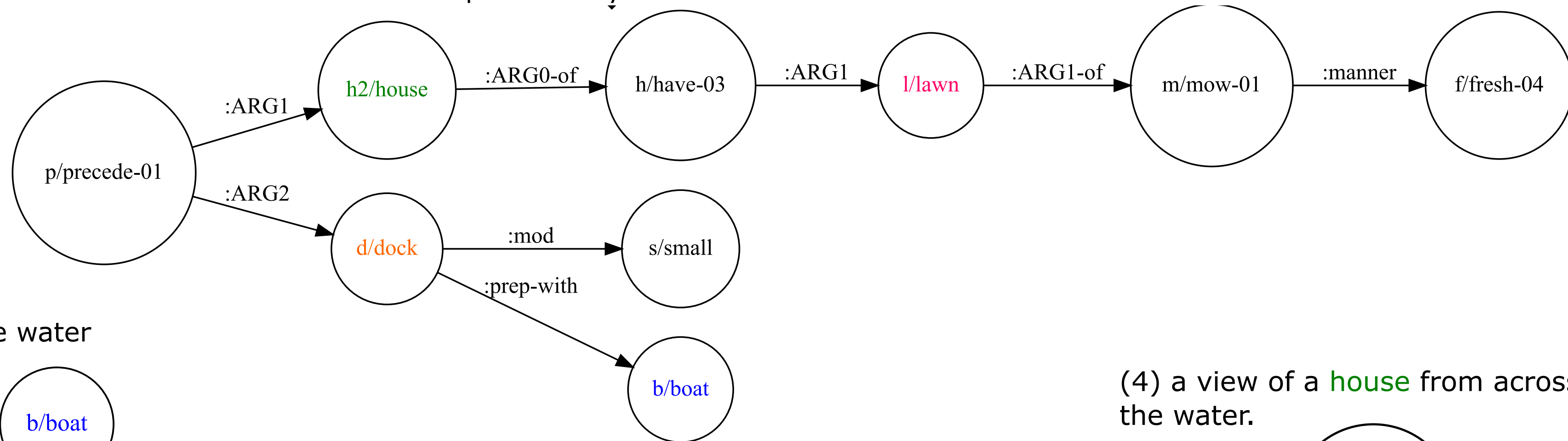
(4) a view of a **house** from across the water.

(5) a large body of water sitting in front of a **house** and green **lawn**.

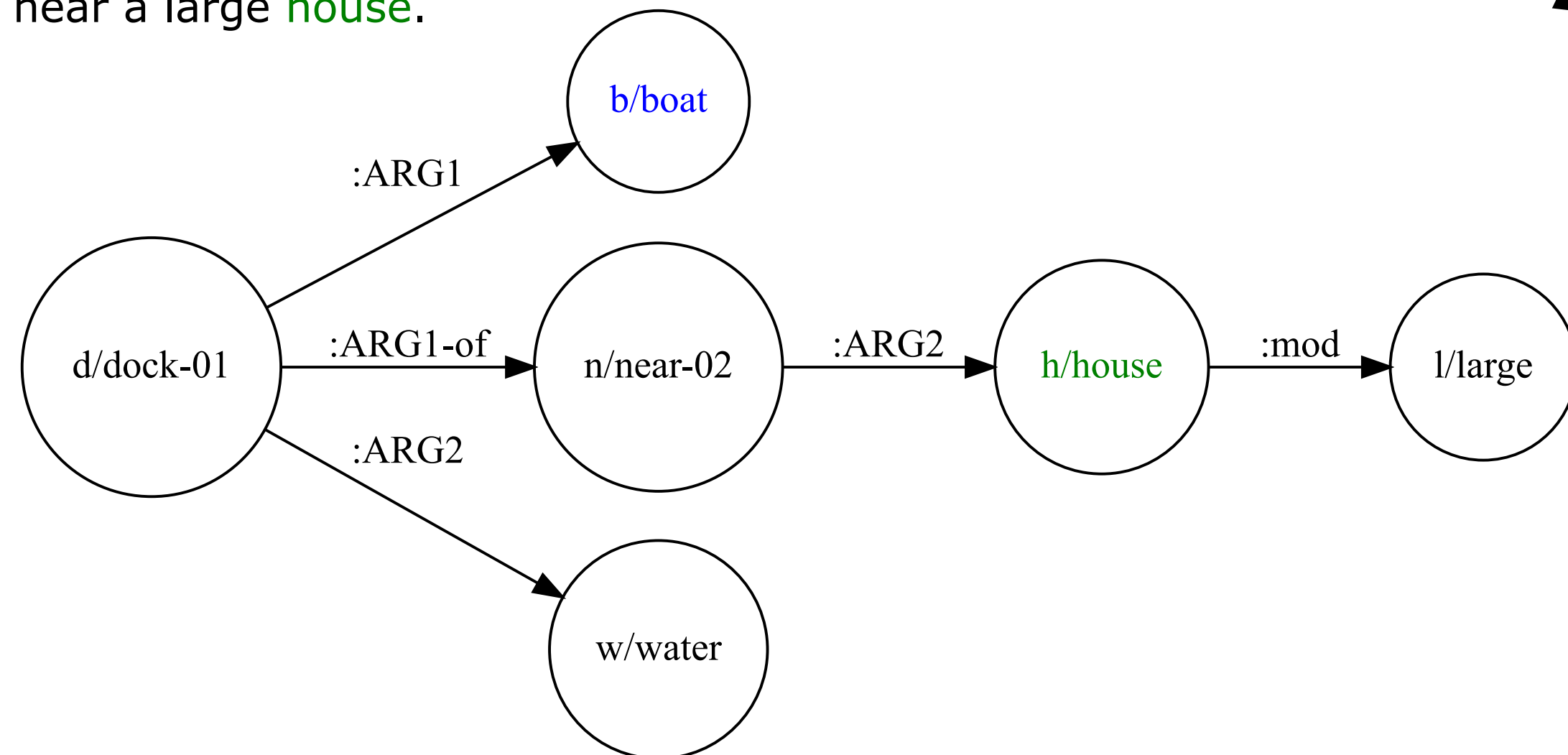
Structured Semantic Augmentations (SSA)

A novel fully-automatic data augmentation approach suitable for CIC.

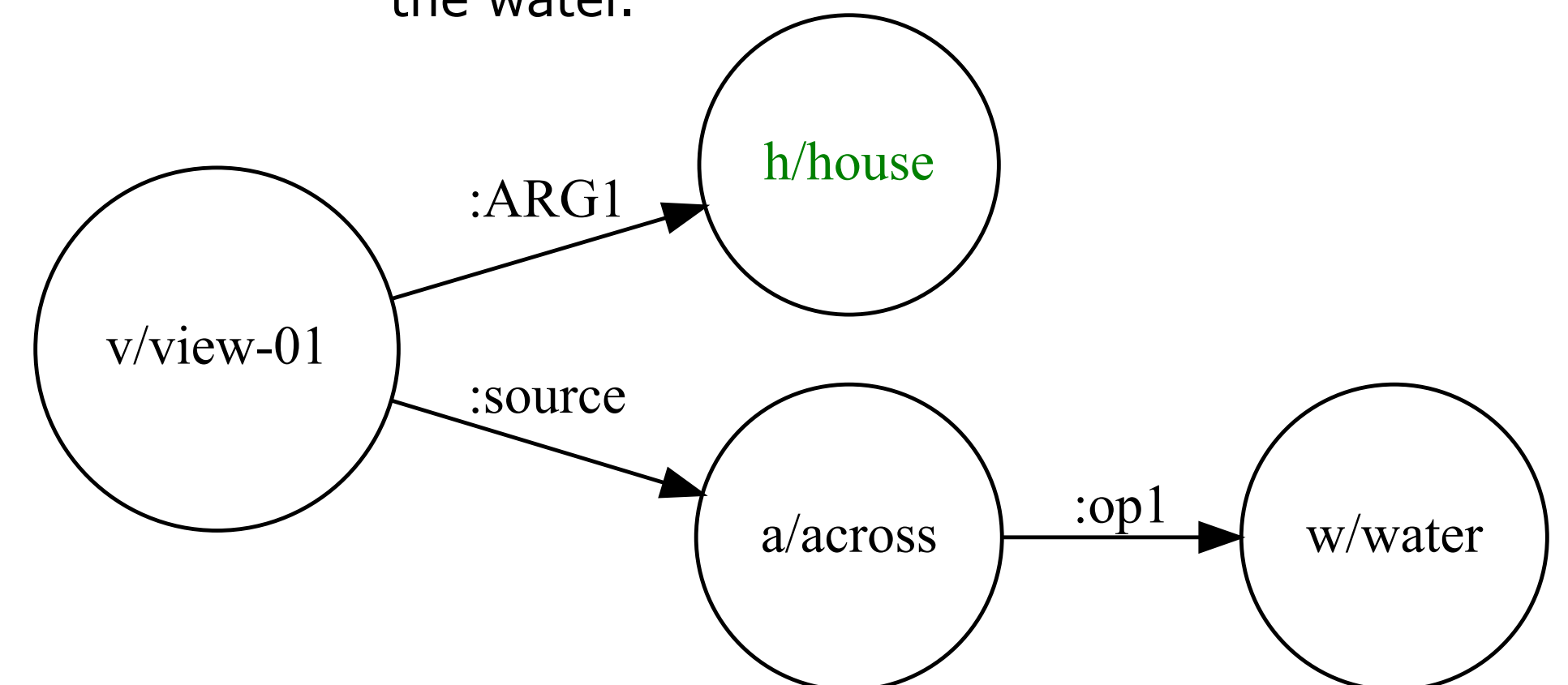
(1) a **house** with a freshly mowed **lawn** is preceded by a small **dock** with a **boat**.



(3) a **boat** is docked in the water near a large **house**.

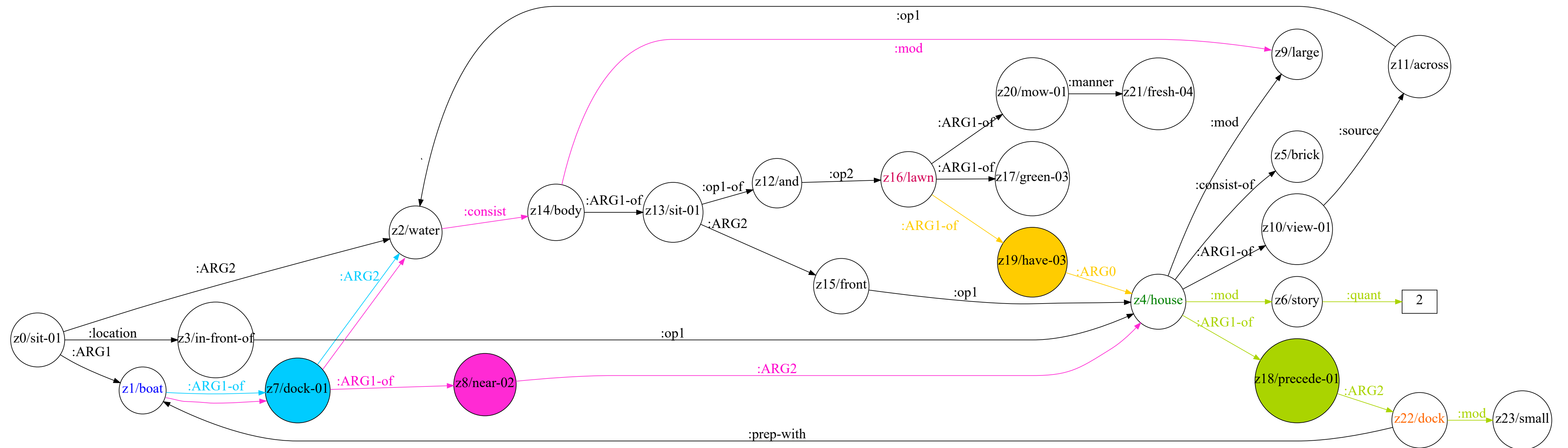


(4) a view of a **house** from across **me** the water.



Structured Semantic Augmentations (SSA)

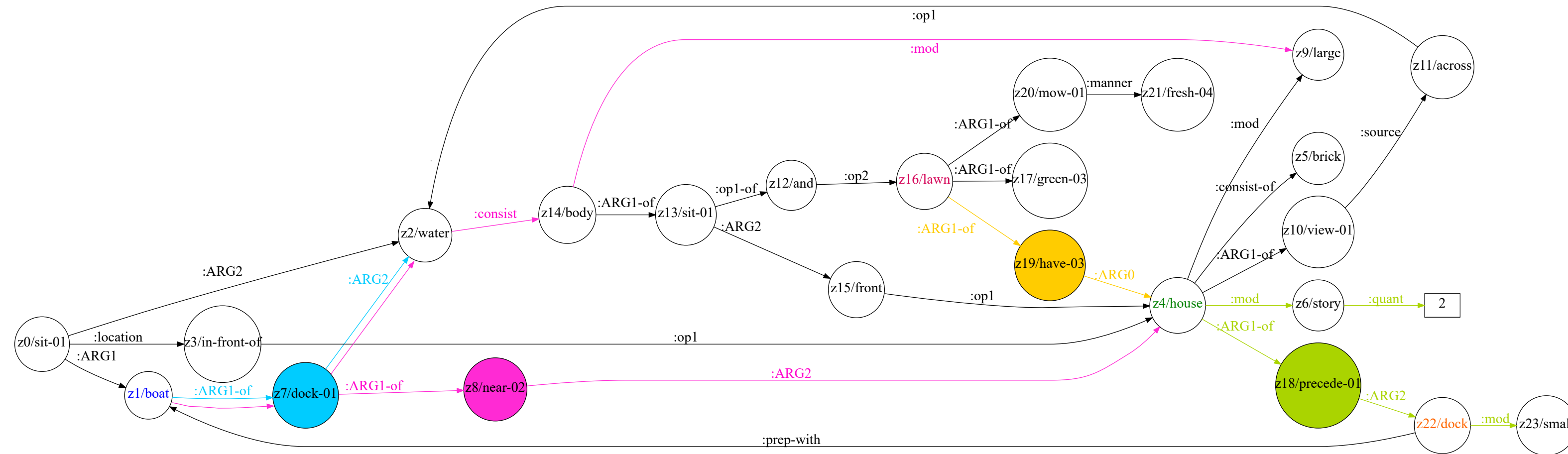
A novel fully-automatic data augmentation approach suitable for CIC.



Meta Visually Grounded AMR (Meta-vgAMR)

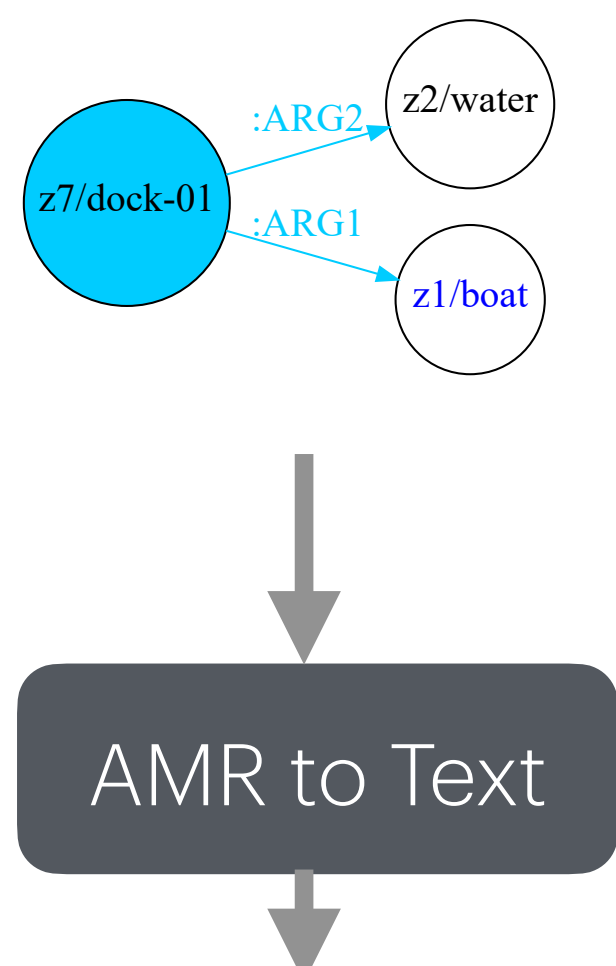
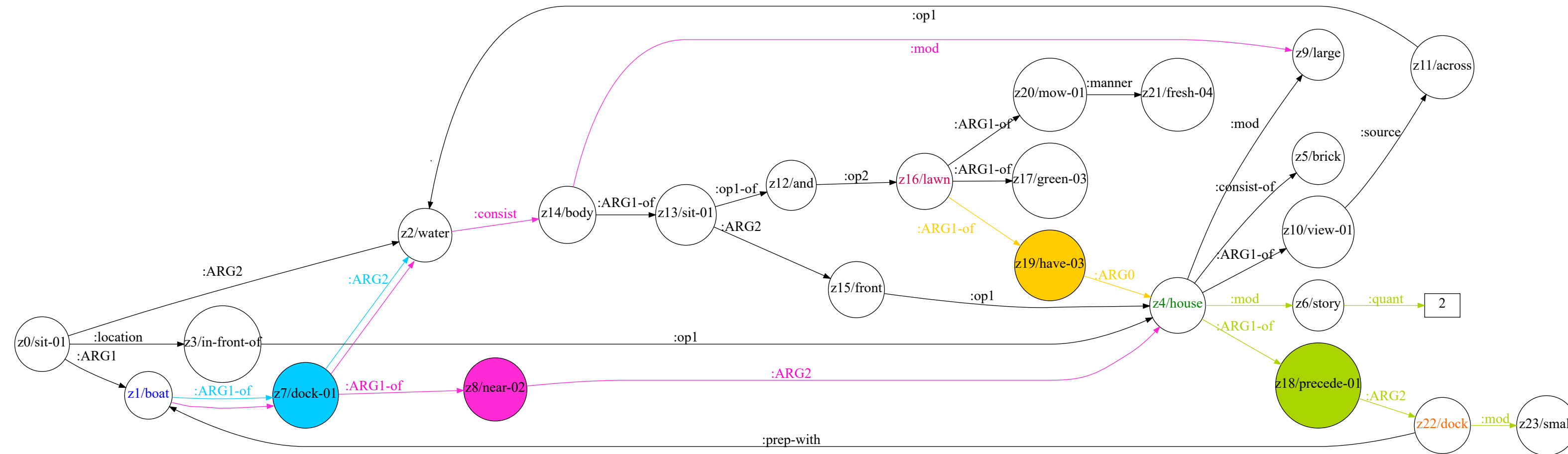
Structured Semantic Augmentations (SSA)

A novel fully-automatic data augmentation approach suitable for CIC.

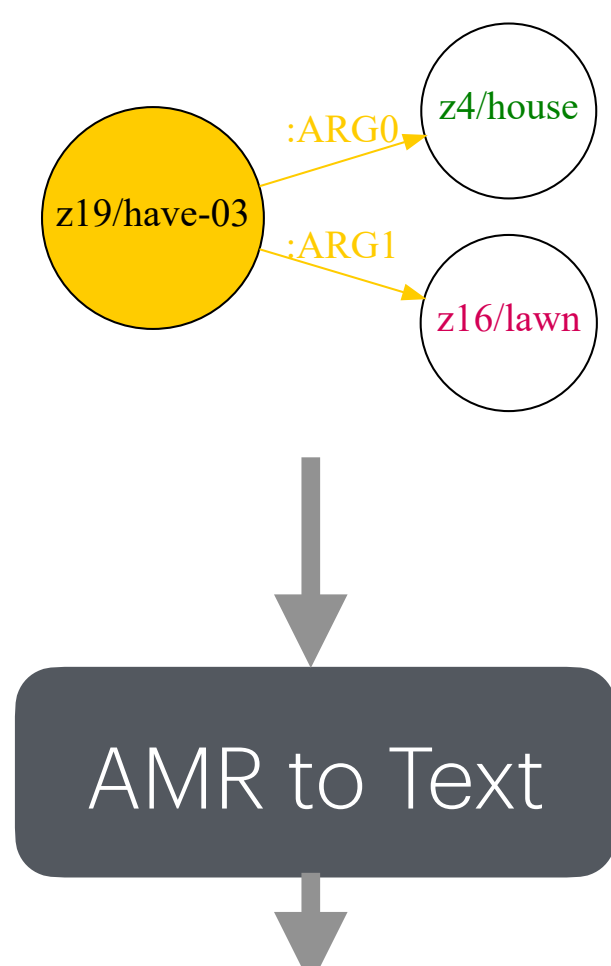


Structured Semantic Augmentations (SSA)

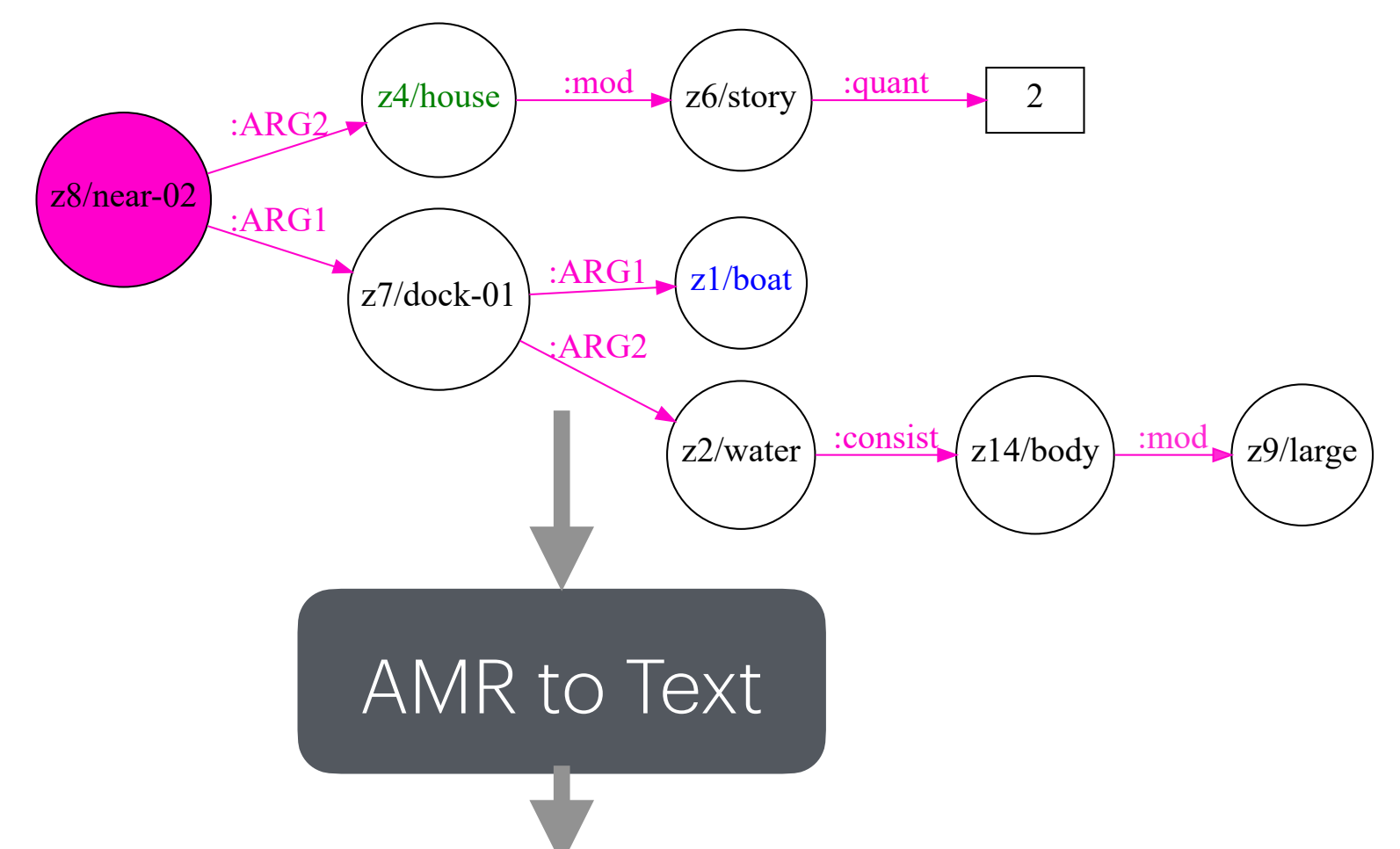
A novel fully-automatic data augmentation approach suitable for CIC.



(a) a boat is docked in the water.



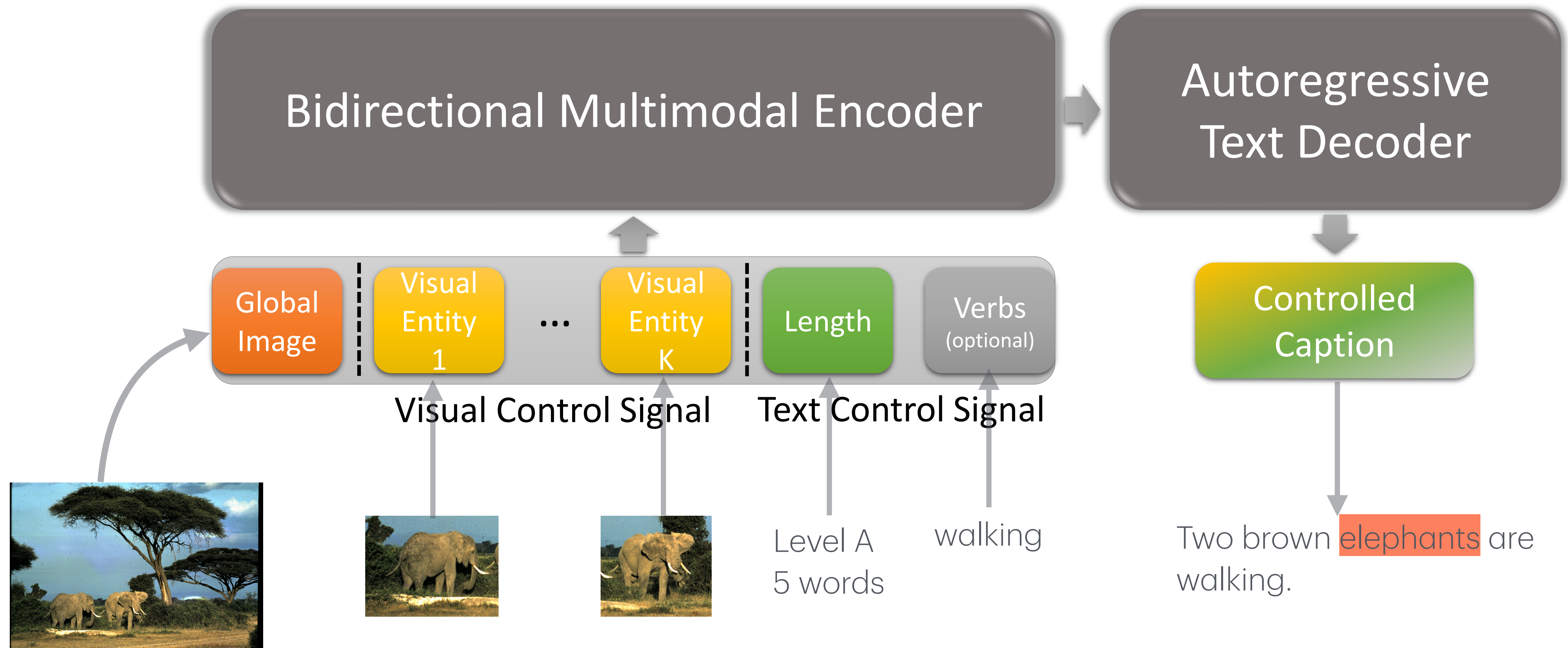
(b) the house has a lawn in front of it.



(e) a boat docked in a large body of water near a two story house.

CIC-BART-SSA

The proposed CIC model.



Quantitative Results

Performance for Original and SSA-only Testing Sets.

Model	H	IoU	G	sC	D-1	D-2	L	H	IoU	G	sC	D-1	D-2	L
	COCO-Ent							Flickr-Ent						
SCT [1]	55.8	67.3	64.4	42.8	27.0	35.5	-	54.6	50.7	79.8	44.0	29.3	36.5	-
ASG [2]	74.2	72.6	72.0	78.3	37.8	<u>56.6</u>	-	-	-	-	-	-	-	-
VSR [3]	56.2	77.6	39.0	67.4	30.0	42.2	-	62.5	60.2	54.0	77.9	33.3	49.3	-
CIC-BART-SSA	78.3	<u>77.2</u>	74.8	82.5	44.6	63.2	0.11	71.3	55.0	86.0	81.7	47.0	62.6	1.05

Metrics

Controllability (IoU, L):

- Content (IoU)
- Length (L)

Text Quality (G):

- GRUEN (G)

Text Diversity (sc,D-1,2):

- Self-CIDEr (sC),
- Distinct n-grams (D-1,2)

Overall CIC performance (H):

- The Harmonic mean (H) of IoU, G, sC.

Model	H	IoU	G	sC	D-1	D-2	L	H	IoU	G	sC	D-1	D-2	L
	COCO-Ent-SSA (SSA only)							Flickr-Ent-SSA (SSA only)						
SCT	51.7	<u>62.1</u>	64.8	37.8	23.7	31.0	-	43.9	29.9	77.3	45.7	31.0	36.7	-
CIC-BART-SSA	75.6	65.2	80.7	83.7	53.8	67.8	0.11	72.0	55.6	82.9	86.1	56.5	69.3	1.05

[1] Cornia, Marcella, et al. "Show, control and tell: A framework for generating controllable and grounded captions." *CVPR*. 2019.

[2] Chen, Shizhe, et al. "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs." *CVPR*. 2020.

[3] Chen, Long, et al. "Human-like controllable image captioning with verb-specific semantic roles." *CVPR*. 2021.

Quantitative Results

Performance for Original and SSA-only Testing Sets.

Model	H	IoU	G	sC	D-1	D-2	L	H	IoU	G	sC	D-1	D-2	L
	COCO-Ent							Flickr-Ent						
SCT [1]	55.8	67.3	64.4	42.8	27.0	35.5	-	54.6	50.7	79.8	44.0	29.3	36.5	-
ASG [2]	74.2	72.6	72.0	78.3	37.8	56.6	-	-	-	-	-	-	-	-
VSR [3]	56.2	77.6	39.0	67.4	30.0	42.2	-	62.5	60.2	54.0	77.9	33.3	49.3	-
CIC-BART-SSA	78.3	<u>77.2</u>	74.8	82.5	44.6	63.2	0.11	71.3	55.0	86.0	81.7	47.0	62.6	1.05

Model	H	IoU	G	sC	D-1	D-2	L	H	IoU	G	sC	D-1	D-2	L
	COCO-Ent-SSA (SSA only)							Flickr-Ent-SSA (SSA only)						
SCT	51.7	<u>62.1</u>	64.8	37.8	23.7	31.0	-	43.9	29.9	77.3	45.7	31.0	36.7	-
CIC-BART-SSA	75.6	65.2	80.7	83.7	53.8	67.8	0.11	72.0	55.6	82.9	86.1	56.5	69.3	1.05

Metrics

Controllability (IoU, L):

- Content (IoU)
- Length (L)

Text Quality (G):

- GRUEN (G)

Text Diversity (sc,D-1,2):

- Self-CIDEr (sC),
- Distinct n-grams (D-1,2)

Overall CIC performance (H):

- The Harmonic mean (H) of IoU, G, sC.

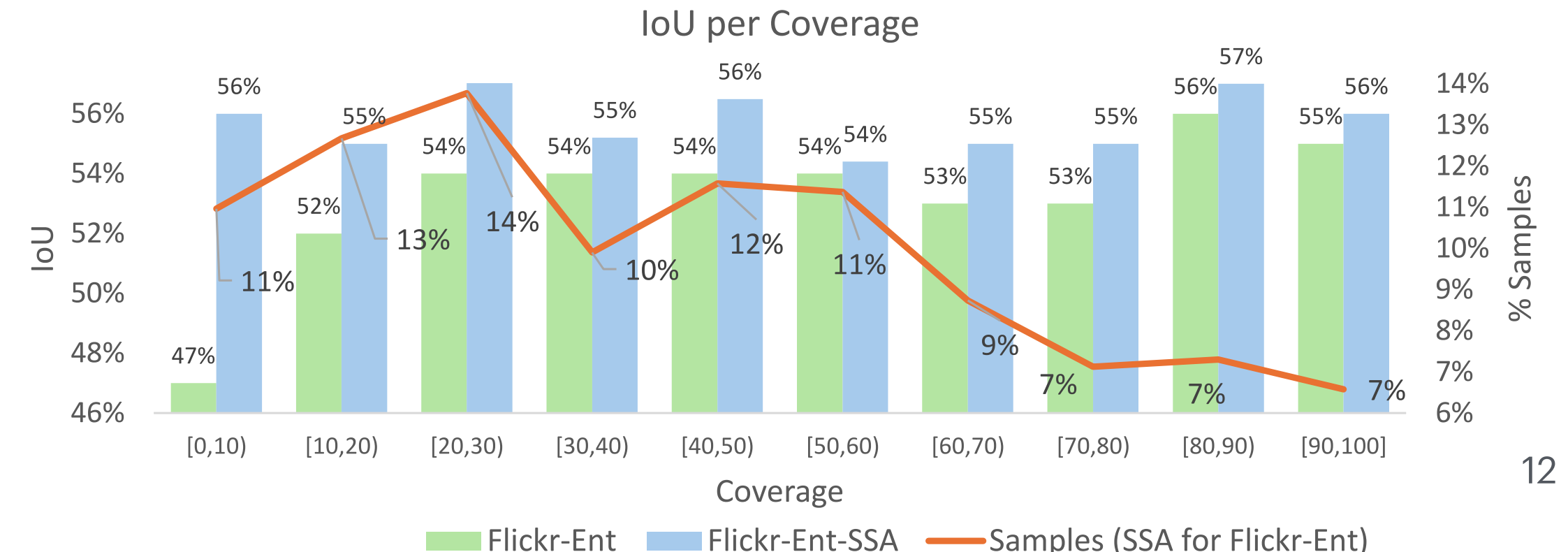
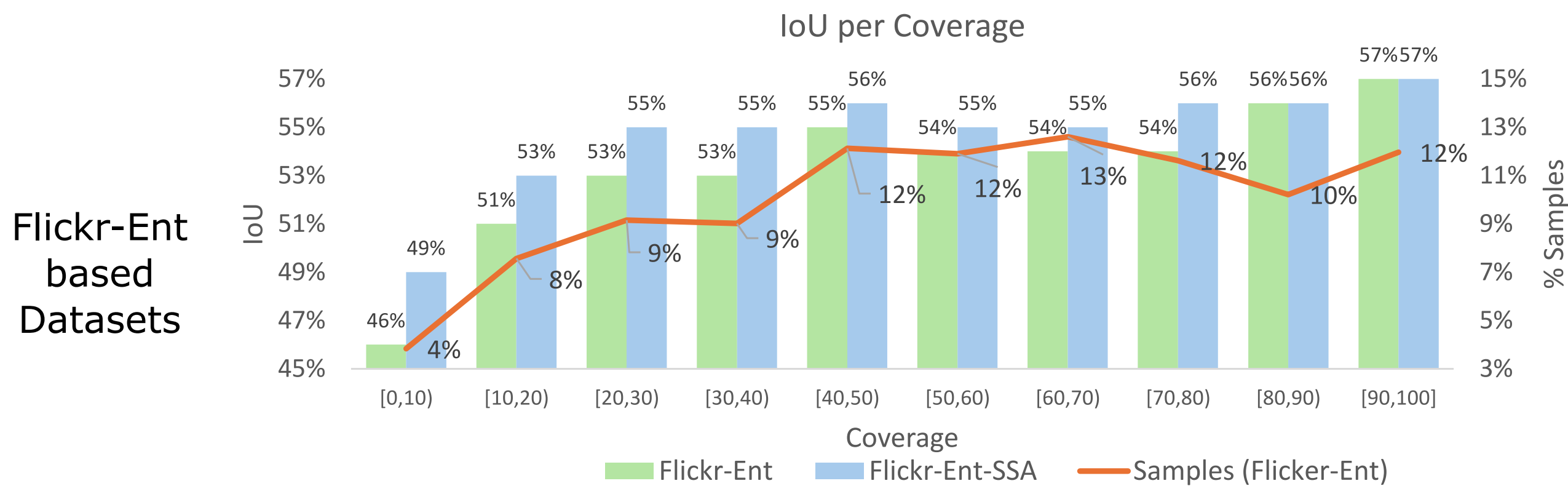
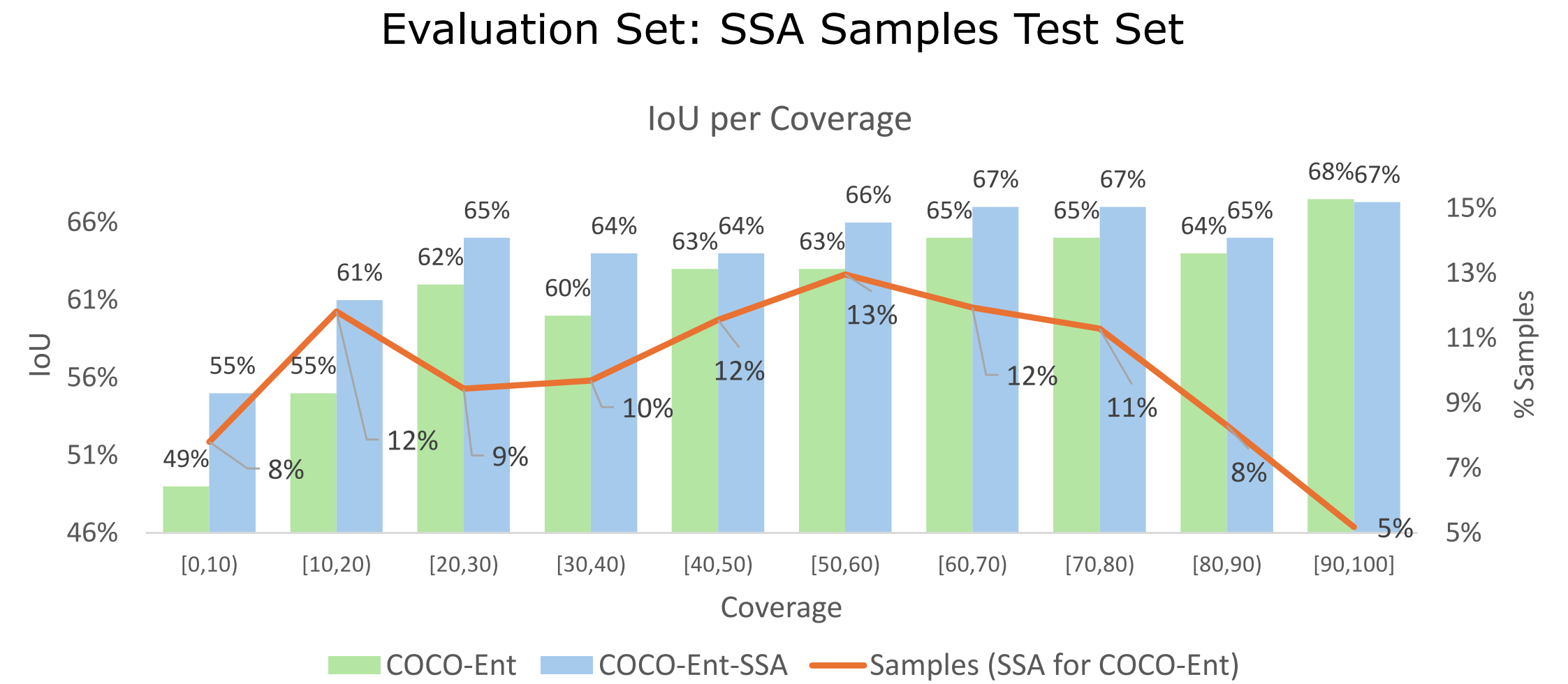
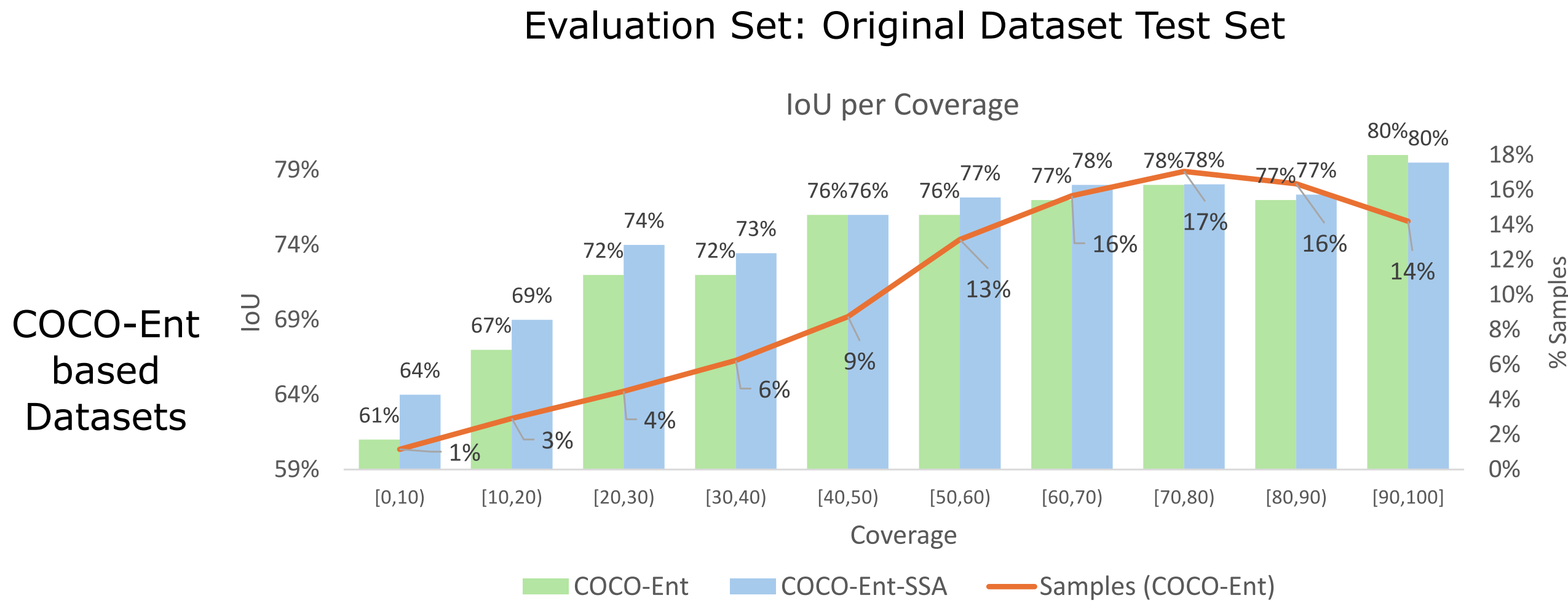
[1] Cornia, Marcella, et al. "Show, control and tell: A framework for generating controllable and grounded captions." *CVPR*. 2019.

[2] Chen, Shizhe, et al. "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs." *CVPR*. 2020.

[3] Chen, Long, et al. "Human-like controllable image captioning with verb-specific semantic roles." *CVPR*. 2021.

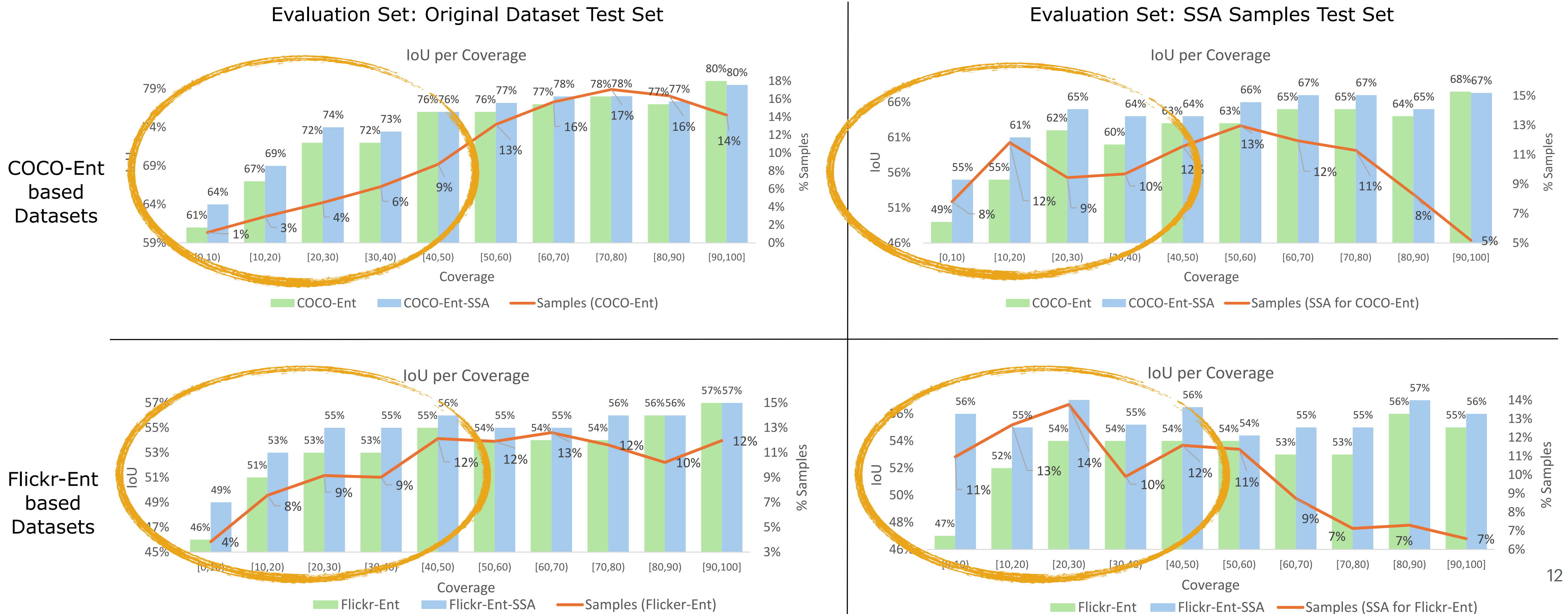
Quantitative Results

Content Controllability (IoU) per coverage band (Coverage=control_signal_area / total_image_area).






Quantitative Results

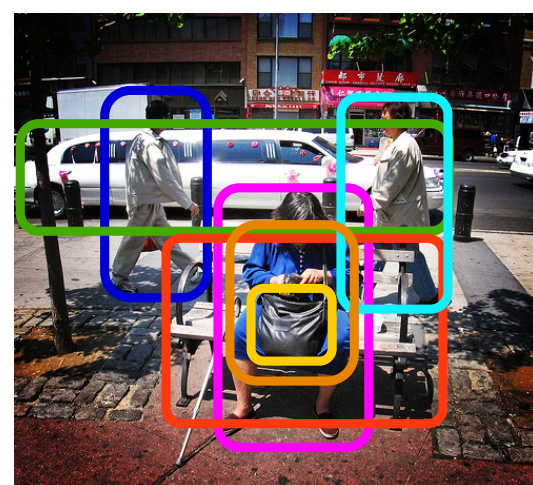
Content Controllability (IoU) per coverage band (Coverage=control_signal_area / total_image_area).



Qualitative Results

Generated Controlled Captions from CIC models.

	SCT	a man cutting a pizza and a pizza
	ASG	a picture of a pizza on a white plate
	VSR	taking a picture of a pizza
	CIC-BART	a man is eating a pizza in a restaurant
	CIC-BART-SSA	a man is taking a picture of a pizza
<hr/>		
	SCT	a man taking a picture of a pizza
	ASG	a man takes a picture of his pizza on a pizza
	VSR	takes a pizza and a pizza on a picture of a man
	CIC-BART	a black and white photo of a man eating a pizza
	CIC-BART-SSA	a man is taking a picture of food at a restaurant
<hr/>		
	SCT	a man sitting at a table with a picture of a pizza and a pizza
	ASG	a man taking a picture of his pizza while sitting at a dinner table
	VSR	taking a picture of a pizza on a table with a man
	CIC-BART	a black and white photo of a man eating at a table
	CIC-BART-SSA	a man is taking a picture of a meal on a table



SCT	a woman in a blue jacket is sitting on a chair on a bench in front of a car
VSR	a woman in a blue jacket sitting on a bench passed a man in a white car
CIC-BART	A woman in a blue dress is sitting on a bench in front of a white car while a man with a briefcase walks by.
CIC-BART-SSA	A woman in a blue dress with a black purse is sitting on a bench in front of a white car as people walk by.

Thank you!

Poster Session Information

- 10:30AM-11:30AM, Wednesday October 2nd
- Poster Session ID: 3
- Poster ID: 98