南京大学
NANJING UNIVERSITY

Bounding and Filling: A Fast and Flexible Framework for Image Captioning

Zheng Ma, Changxin Wang, Bo Huang, Zixuan Zhu and Jianbing Zhang

Speaker: Changxin Wang

Agenda

- **Introduction**
  - Definition for Image captioning
  - Generation manners and problems
- **Method**
  - Proposal
  - Details: architecture, training and inference
- **Results**
- **Conclusions**

Introduction

Definition for Image Captioning

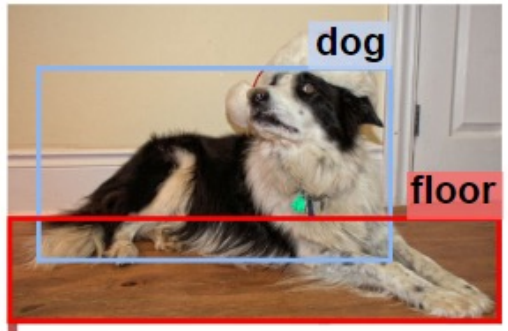
Image Captioning

A cute dog lying on the floor

Image Captioning: generating a sentence to describe given image

Visual understanding + Text generation

Generation Manners



(a) *Autoregressive Mode*

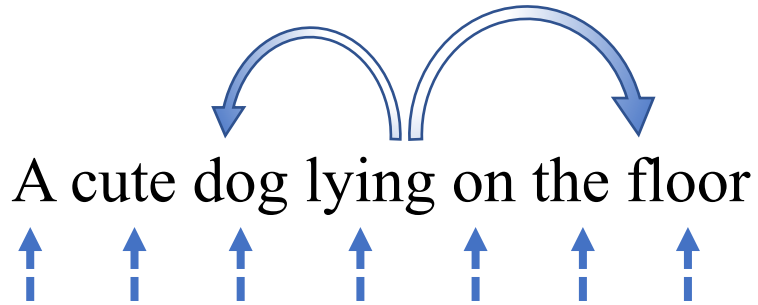
→ a → cute → dog → lying → on → the → floor

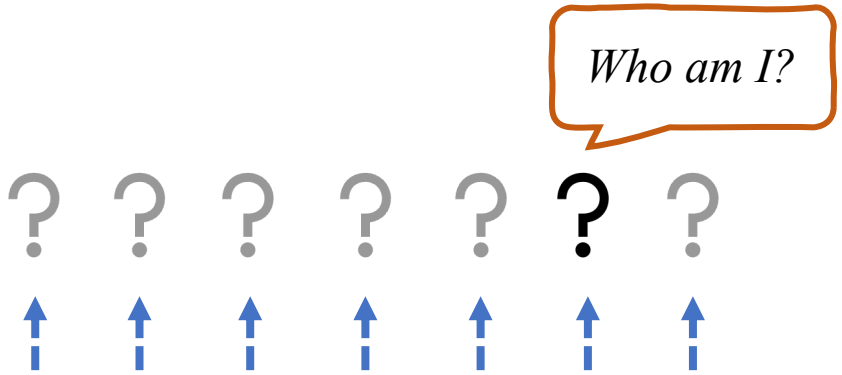
(b) *Non-Autoregressive Mode*

→ a cute dog lying on the floor

(c) *Semi-Autoregressive Mode*

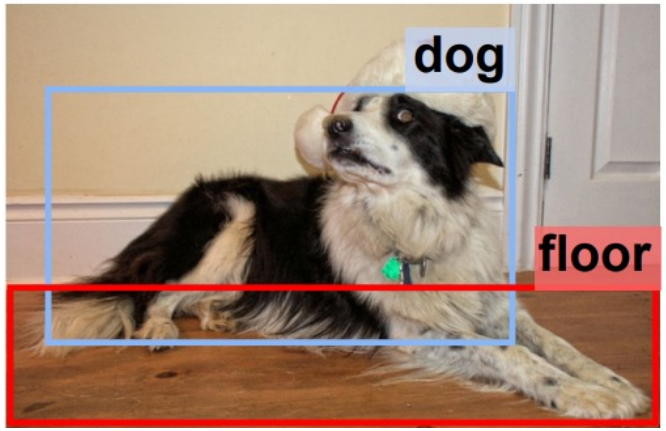
→ a cute → dog lying → on the → floor

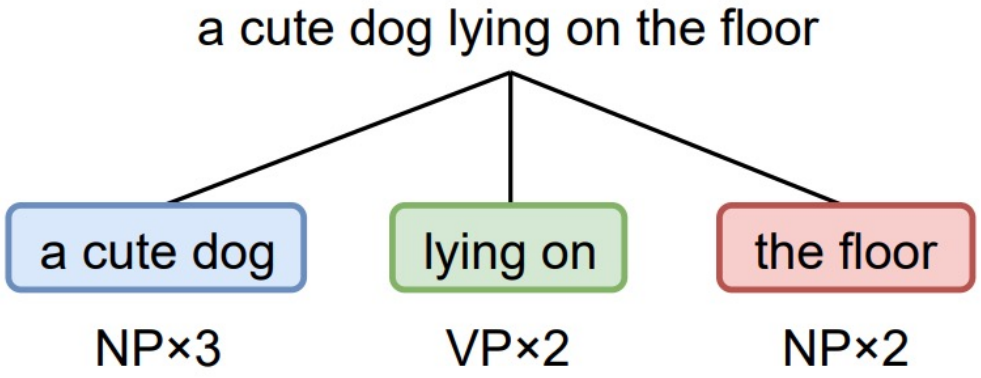




Method

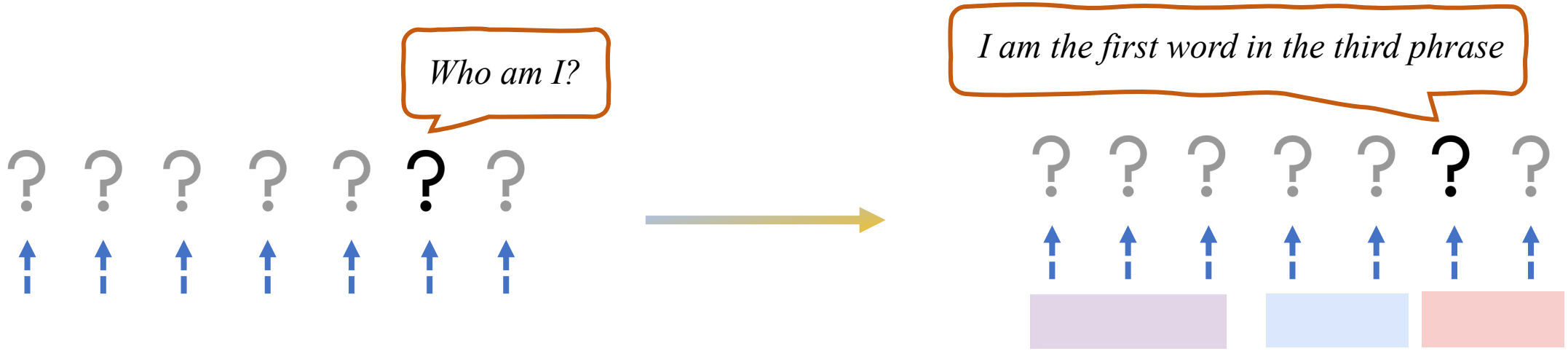
Characteristic of Image Captioning



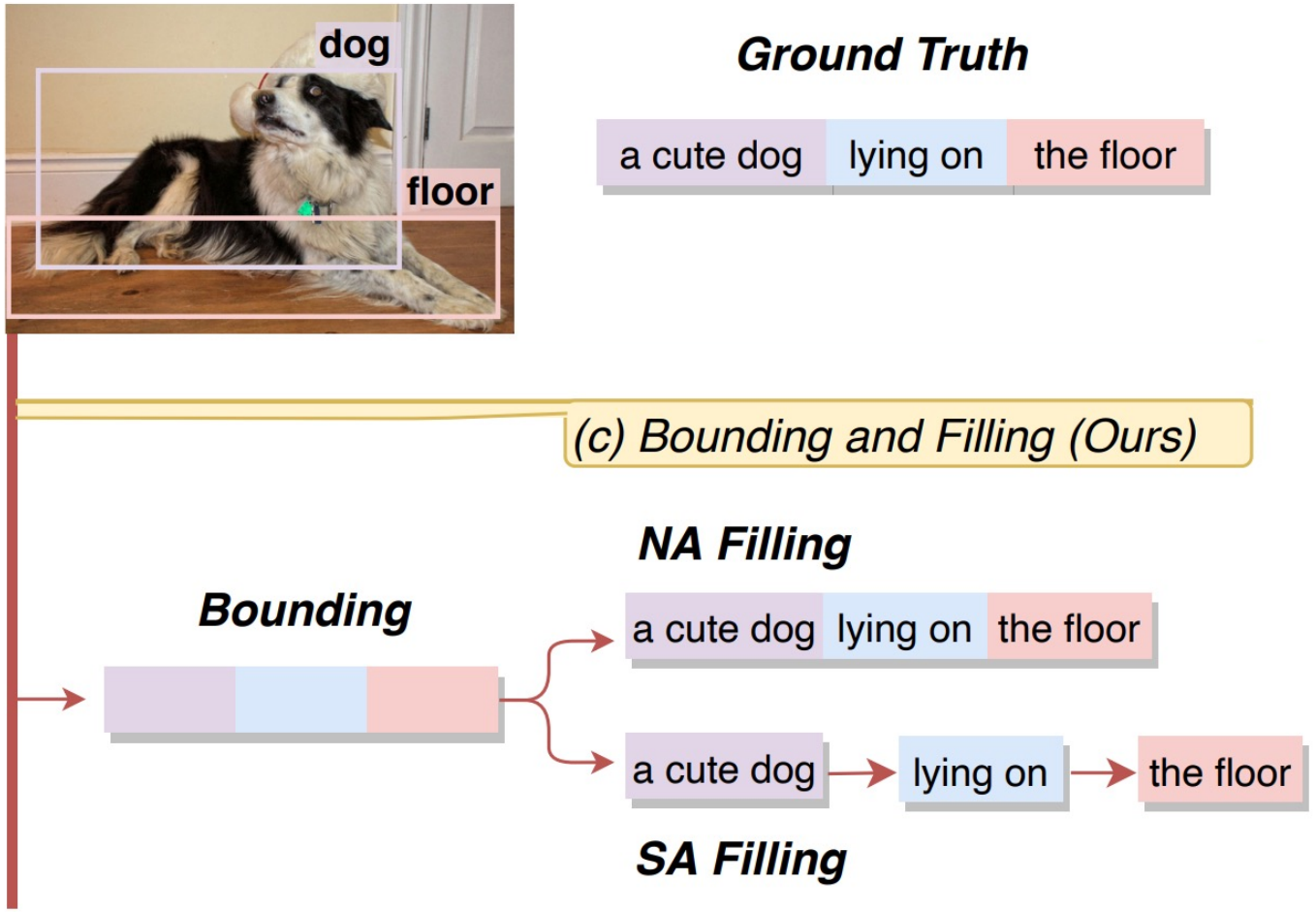


A description typically consists of meaningful phrases

Where phrase info can be used as hint to facilitate generation

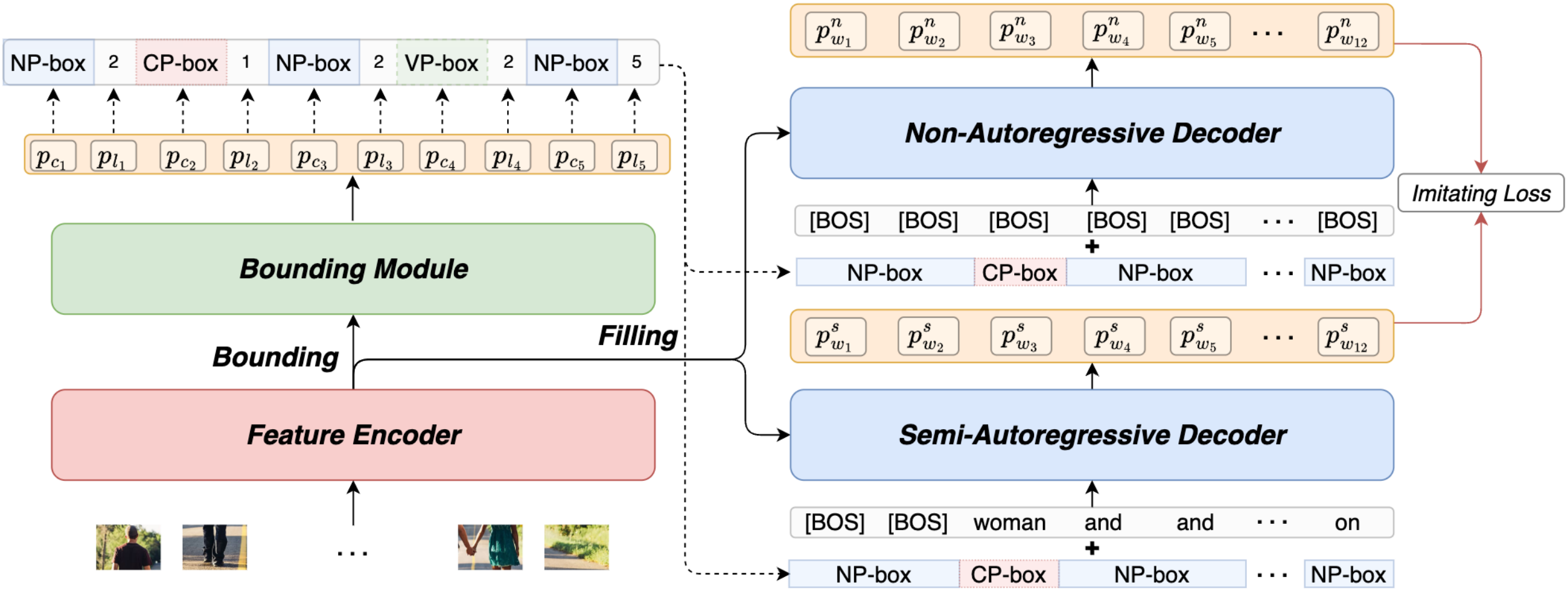


Bounding and Filling



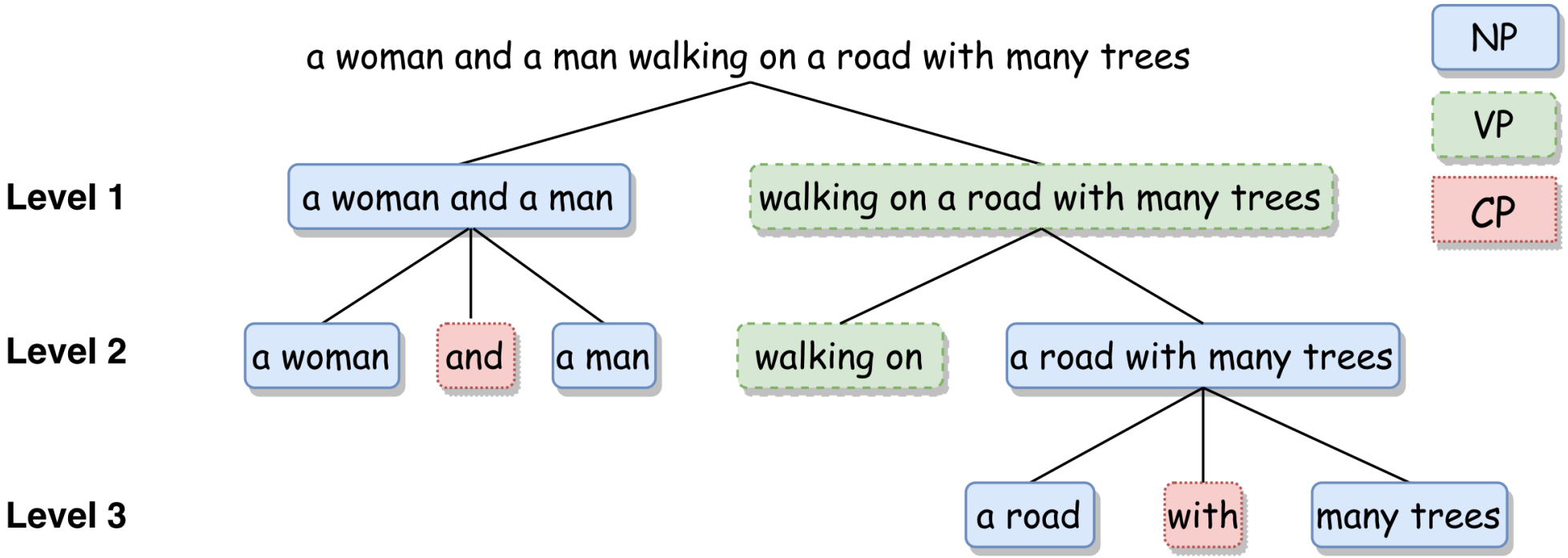
Bounding: predict a series of bounding boxes to facilitate description generation
Filling: populate the boxes with the appropriate words

Architecture



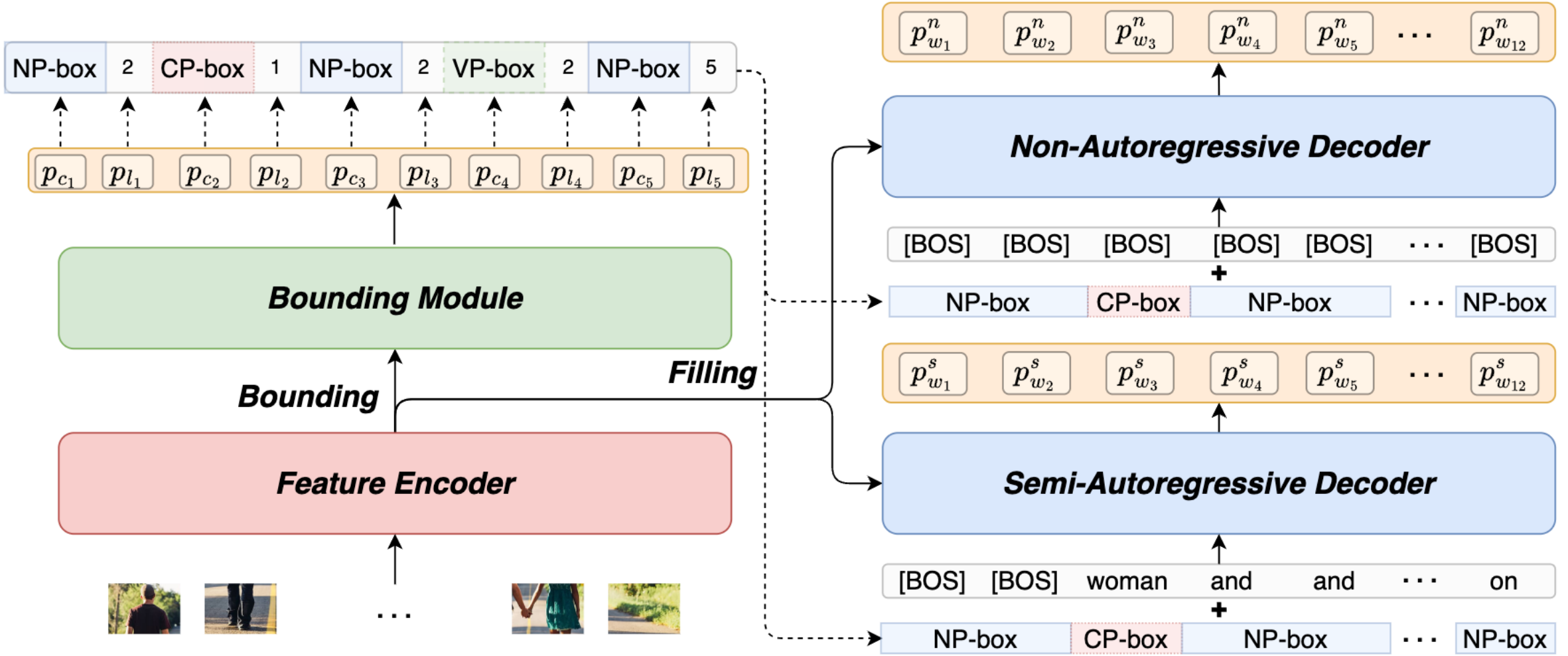
Bounding Module + Shared Decoder + Imitate Strategy

Hierarchical Boxes



A description is parsed into a tree structure by a constituency parser

Inference



Bounding: predict a series of bounding boxes autoregressively

Filling: populate the boxes with the appropriate words in two manners

Results

Main Comparison Results

Models	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr	SPICE	Latency	Speedup
Autoregressive Models								
AIC (bw=1)	80.5	38.9	29.0	58.7	129.4	**22.8**	192ms	**1.73×**
AIC (bw=3)	**80.9**	**39.3**	29.0	**58.9**	130.2	**22.8**	332ms	1.00×
Non-Autoregressive Models								
MNIC [9]	75.4	30.9	27.5	55.6	108.1	21.0	-	2.80×
FNIC [6]	/	36.2	27.1	55.3	115.7	20.2	-	8.15×
CMAL [10]	**80.3**	37.3	28.1	58.0	124.0	21.8	-	**13.90×**
IBM [7]	77.2	36.6	27.8	56.2	113.2	20.9	-	3.06×
BoFiCap-NA (*Ours*)	80.1	**38.2**	**28.4**	**58.2**	**125.6**	**22.1**	36ms	9.22×
Semi-Autoregressive Models								
PNAIC [8]	80.4	38.3	29.0	58.4	**129.4**	22.2	-	2.17×
SATIC [31]	**80.8**	38.4	28.8	58.5	129.0	**22.7**	-	1.65×
SAIC [28]	80.4	38.7	**29.4**	58.5	128.3	22.2	-	1.55×
BoFiCap-SA (*Ours*)	80.5	**38.9**	28.8	**58.8**	128.4	**22.7**	90ms	**3.69×**

Performance comparisons with different evaluation metrics on the test set of MS COCO
BoFiCap achieve sota performance under non-autoregressive manner

Ablation Study

Models	Methods	B4	M	R	C	S
NAIC	/	30.8	25.6	55.4	108.5	19.6
BoFiCap-NA	+Bound	37.6	28.1	58.0	122.3	21.8
	+Bound+Joint	37.9	28.3	58.2	124.7	22.0
	+Bound+Joint+Imit	38.2	28.4	58.2	125.6	22.1
BoFiCap-SA	+Bound	38.9	28.8	58.8	128.4	22.7
	+Bound+Joint	38.6	28.7	58.5	127.2	22.7
	+Bound+Joint+Imit	38.5	28.7	58.5	127.5	22.6

Effect of three methods evaluated on MS COCO test set, where Bound, Joint, and Imit represent the bounding boxes, shared parameters, and imitating strategy.

Case Study

AIC: a black and white photo of street signs in front of a building.

NAIC: a black and white photo of street street with street signs in front of a building.

BoFiCap-NA: a black and white photo of street building with a clock tower.

BoFiCap-SA: a black and white photo of street signs with a clock tower.

GT : an old black and white photo of pennsylvania avenue.

AIC: a black and white cow standing in a field.

NAIC: a cow standing on on grass in the field.

BoFiCap-NA: a black cow standing in the grass in a field.

BoFiCap-SA: a black cow standing in the grass in a field.

GT : a cow stands in the grassy area of a yard.

BoFiCap-NA: a donut and a cup of coffee on a table.
(given boxes: NP×2 CP×1 NP×4 CP×1 NP×2)

BoFiCap-NA: a donut sitting on a wooden table.
(given boxes: NP×2 VP×2 NP×3)

BoFiCap-NA: a cup of coffee next to a donut.
(given boxes: NP×4 CP×2 NP×2)

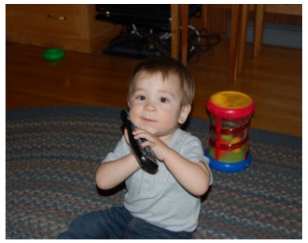
GT : a cup of coffee and a doughnut are on a table.

BoFiCap-SA: a man and a woman standing next to each other.
(given boxes: NP×2 CP×1 NP×2 VP×3 NP×2)

BoFiCap-SA: a man with a tie and a woman in a building.
(given boxes: NP×2 CP×1 NP×2 CP×1 NP×2 CP×1 NP×2)

BoFiCap-SA: a couple standing in a building.
(given boxes: NP×2 VP×2 NP×2)

GT : a man standing next to a woman inside of a building.



BoFiCap-NA (k= 1): a young boy sitting on the floor holding a phone.

BoFiCap-NA (k= 2): a young boy sitting on the floor holding a cell phone.

BoFiCap-NA (k=-1): a young boy sitting on the floor holding a cell phone.

GT : a young boy sitting on a rug holding a cell phone.

BoFiCap-SA (k= 1): a man sitting on a couch with a cat and a laptop.

BoFiCap-SA (k= 2): a man sitting on a couch with a cat and a laptop.

BoFiCap-SA (k=-1): a man sitting on a couch with a cat and a laptop.

GT : a man sitting in a chair with a cat and a laptop.

Top line: **BoFiCap can generate fluent and precise description under both NA and SA manners**

Middle line: **BoFiCap can generate diverse description according to user specified bounding boxes**

Conclusions

Conclusions

- **Bounding and Filling instead of vanilla generation**
- **Fast:**
  - Non-autoregressive manner
  - Semi-autoregressive manner
- **Flexible:**
  - Training once, switch decoding manner according to user requirements
  - Generating diverse description according to user specified bounding boxes
- **Sota performance in non-autoregressive manner**

Thanks
Q&A