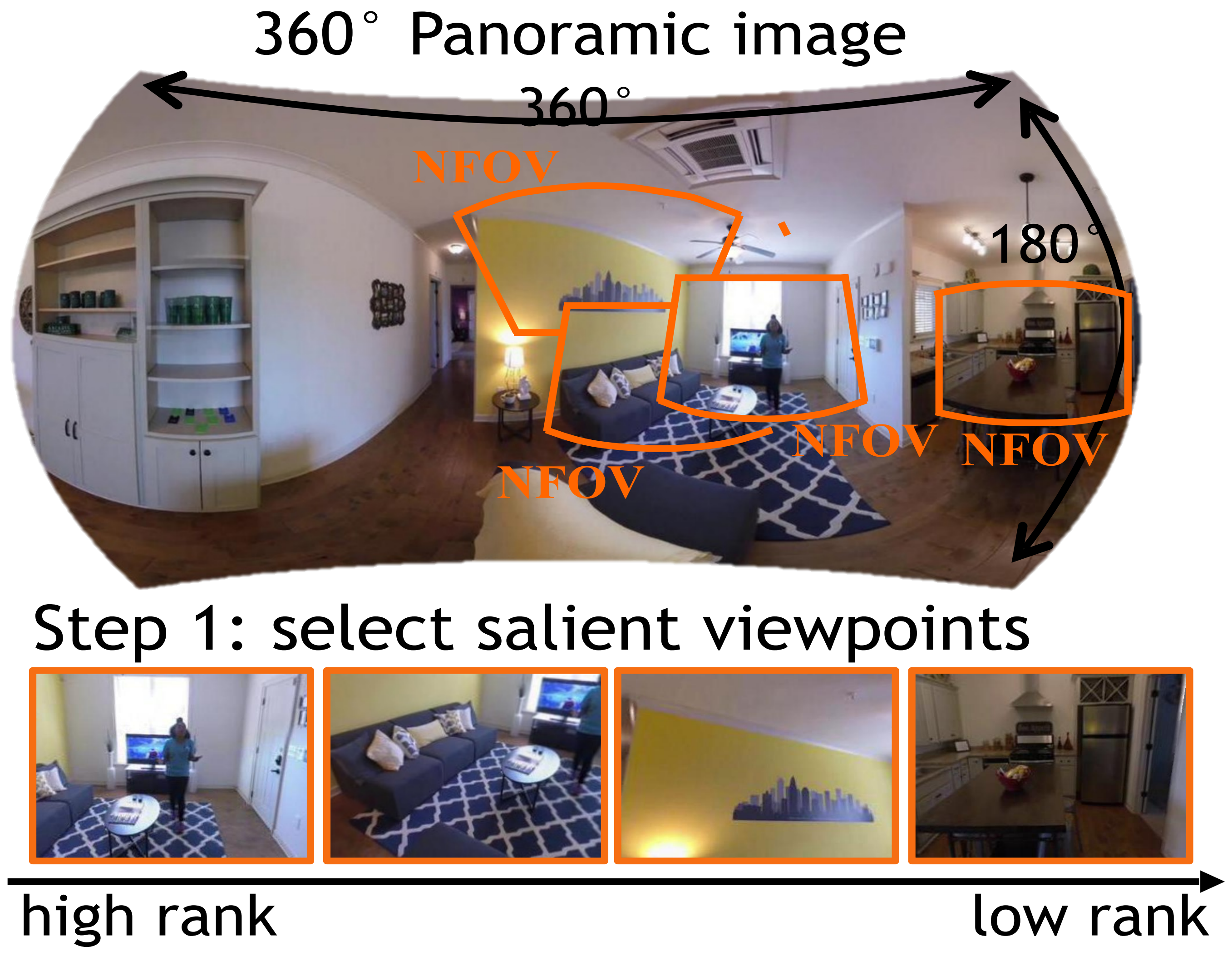
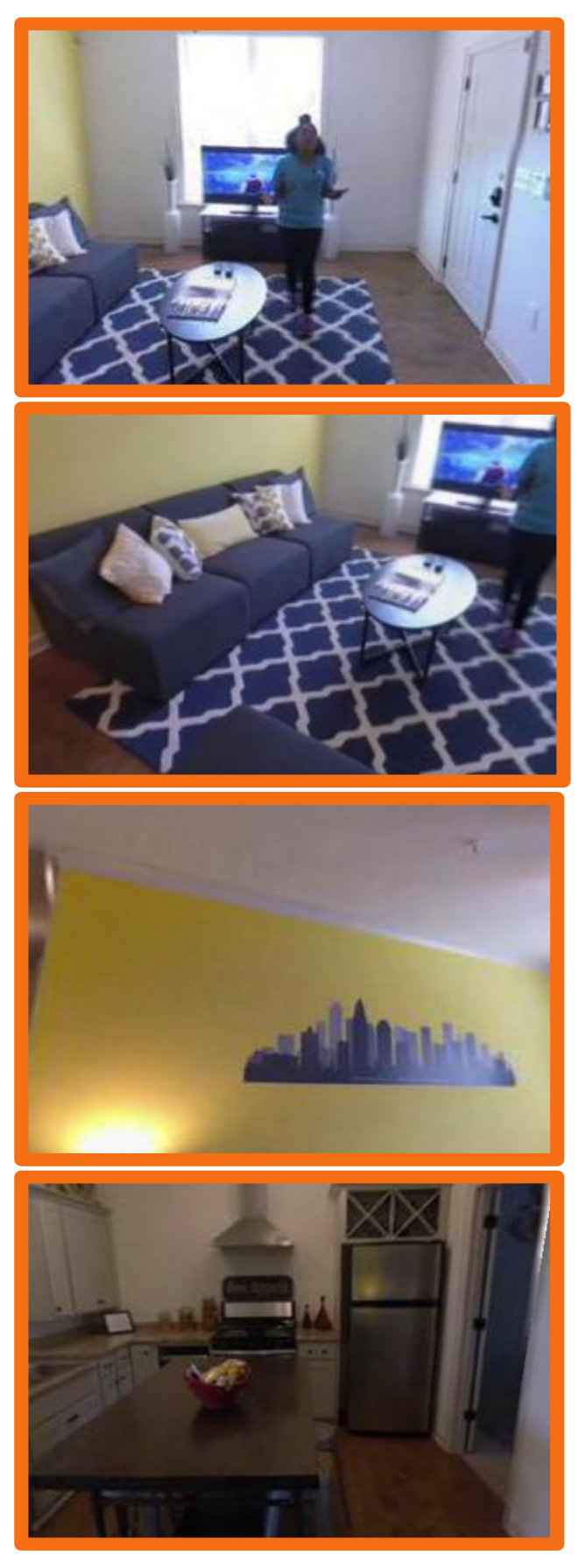
Towards 360° Show-and-Tell

Shih-Han Chou[†], Yi-Chun Chen[†], Cheng Sun[†], Kuo-Hao Zeng[†], Ching Ju Cheng[†], Jianlong Fu[‡], Min Sun[†]
[†]National Tsing Hua University, Hsinchu, Taiwan, [‡]Microsoft Research, Beijing, China

PROBLEM FORMULATION



Step 2: natural language description generation



We are inside of the common area right now.

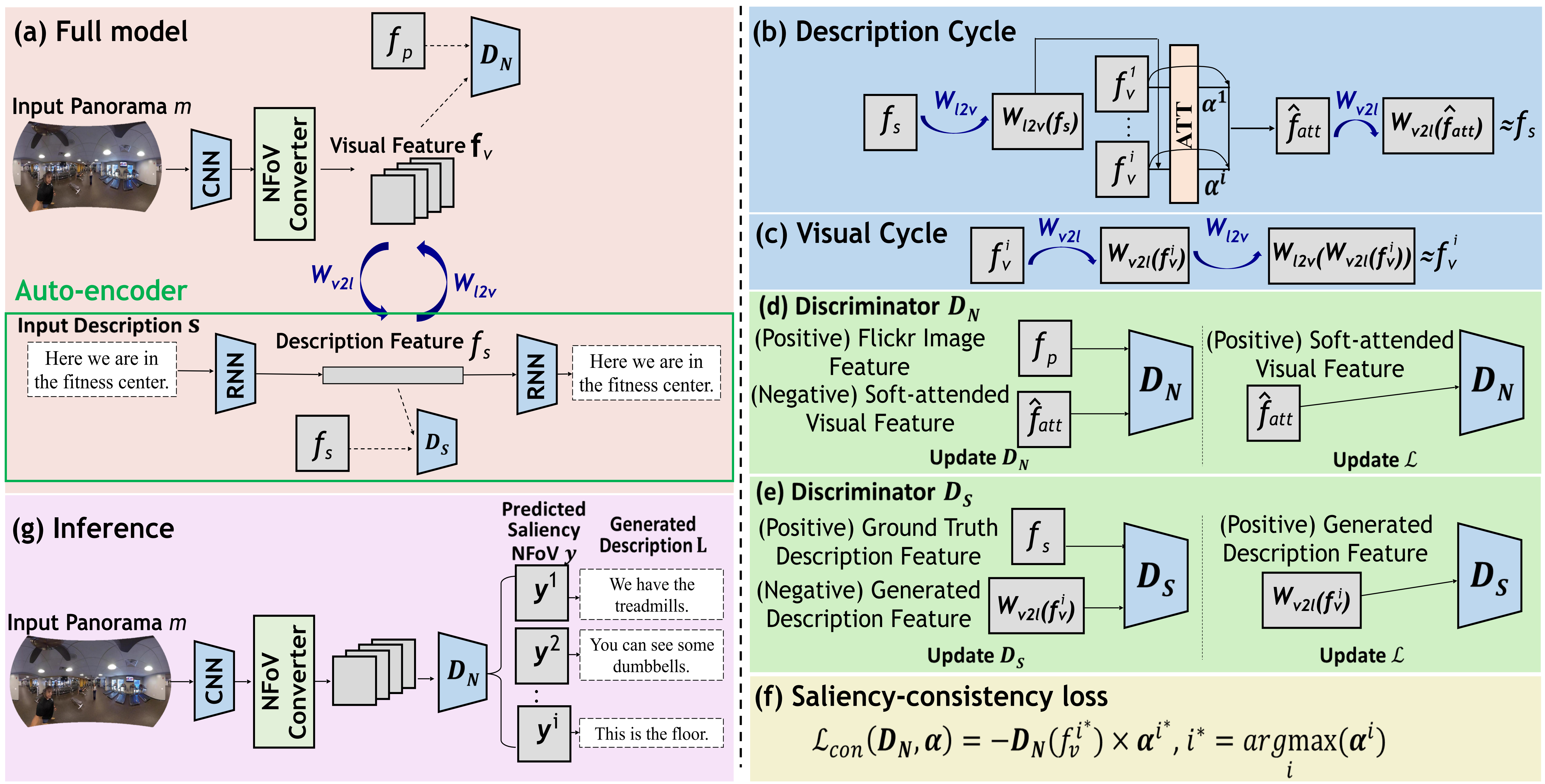
As you can see we have the couch.

You are also able to paint your walls.

Here you also have the kitchen appliances.

Our proposed show-and-tell model can first select the saliency viewpoints (step 1) and generate natural language descriptions (step 2) in the 360 images.

PROPOSED MODEL



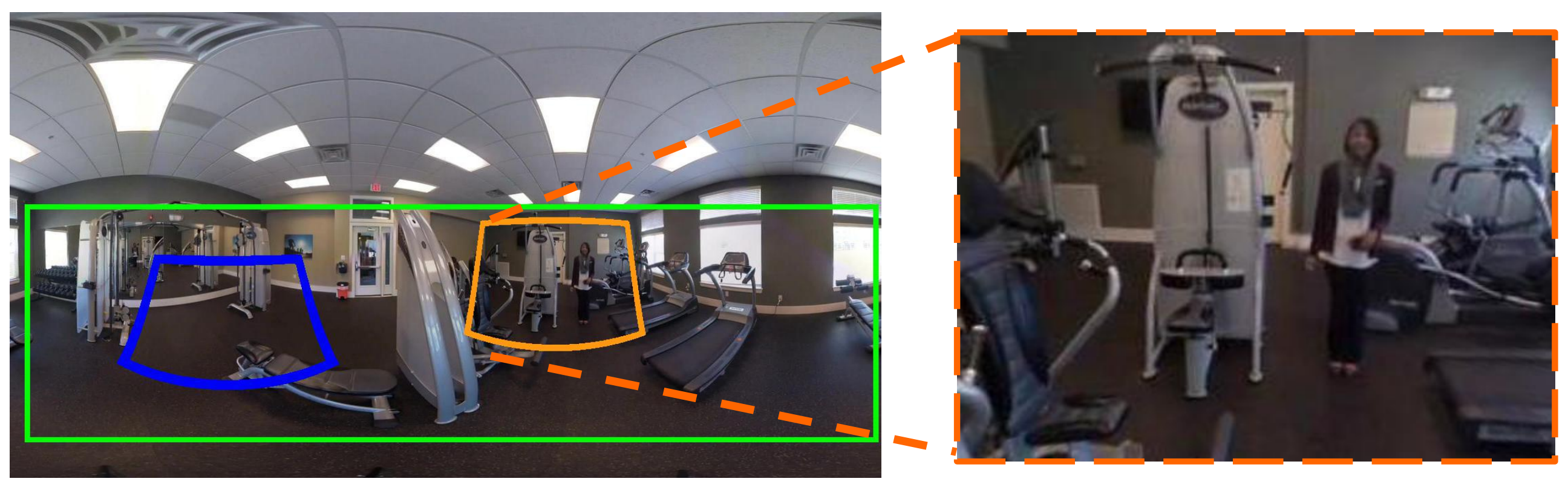
$$\mathcal{L}_{con}(D_N, \alpha) = -D_N(f_v^{i^*}) \times \alpha^{i^*}, i^* = \underset{i}{argmax}(\alpha^i)$$

(a) Training: Panorama/descriptions pass through the Encoder.
(b) Description cycle: Soft-attention is applied for visual grounding.
(c) Visual cycle: NFoV candidates must be consistent with original features.
(d)&(e) Discriminators: Saliency examples are collected to adjust the model to make two discriminators unable to differentiate the positive and negative samples.
(f) Saliency-consistency loss: Ensure descriptions are grounded in saliency NFoVs.
(g) Inference: The model takes panorama as input to rank the NFoV candidates and generate the descriptions.

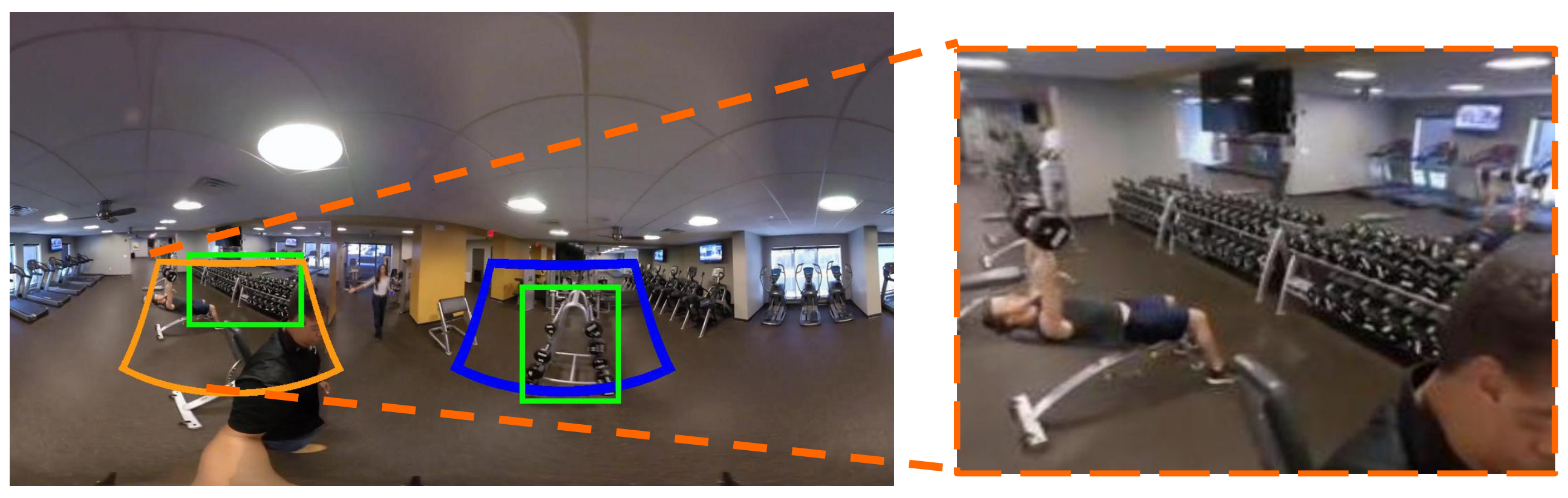
SHOW AND TELL CAPTIONING RESULT

Approach	B@1	B@2	B@3	B@4	ROUGE-L	METEOR	CIDEr
Visual-attention	28.3	12.6	5.1	1.4	30.8	8.5	38.6
Regions-Hierarchical	30.8	18.4	8.7	**4.2**	34.2	7.0	49.4
Ours	**37.2**	**22.1**	**9.3**	0.1	**37.1**	**9.7**	**54.4**

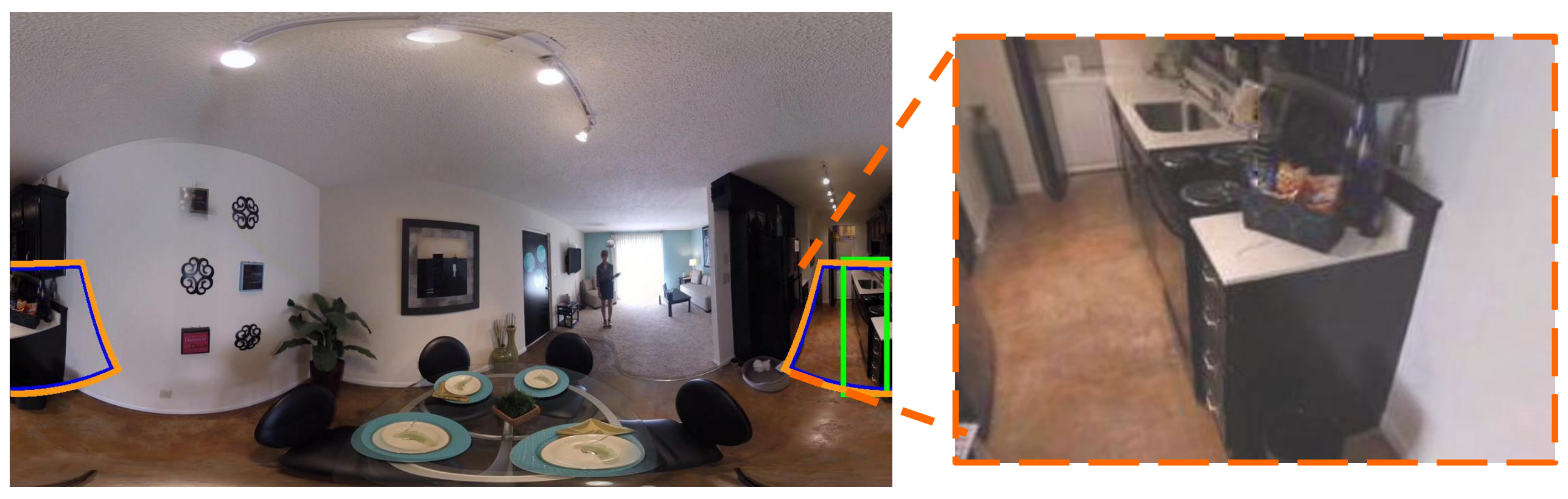
NFoV GROUNDING RESULT



Welcome to our gym.



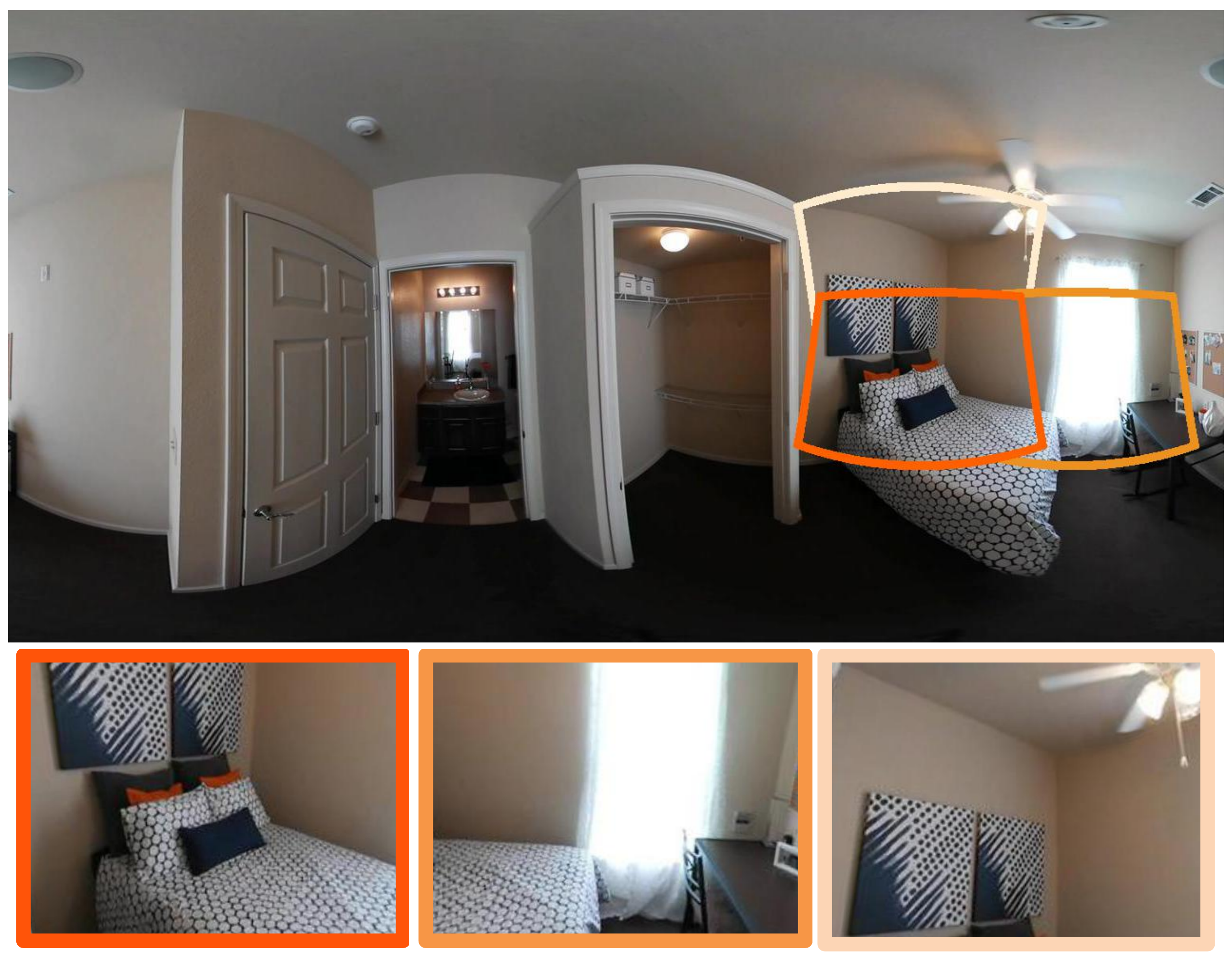
You will have the free weights.



You can see the countertop here.

Orange: Predicted NFOV Green: Annotated Bounding Box
Blue: NFOV Ground-truth

SHOW AND TELL RESULT



Here will have the mattress.

You will see the chairs.

You can see the painting.

VISUAL GROUNDING RESULT

Approach	Avg. Recall	Avg. Precision
VGM	12.3	21.2
Ours	**16.0**	**25.5**
Oracle	62.7	60.9

VIEWPOINT RANKING RESULT

Approach	Rank-1	Rank-2
RR	2.34	3.25
Ours	**2.12**	**2.80**
Oracle	1.00	1.00