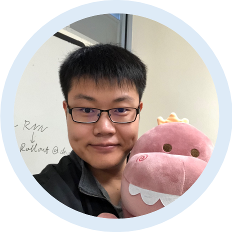
ICML
International Conference On Machine Learning

3D Question Answering via only 2D Vision-Language Models

Fengyun Wang[1] Sicheng Yu[2] Jiawei Wu[3] Jinhui Tang[4] Hanwang Zhang[1] Qianru Sun[2]

1 NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

2 SMU SINGAPORE MANAGEMENT UNIVERSITY

3 NUS National University of Singapore

4 NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

3D Question Answering

answer natural language questions based on 3D scenes

Inputs:

<Question>

What is the black couch facing?

<3D Scene>

Outputs:

<Answer>:

coffee table

1 Why Use 2D LVLM: Lack of Large-scale 3D-language Data

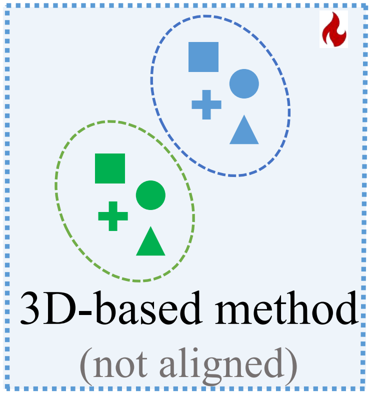
Strategies for achieving effective 3D-Language alignment

Inputs:

<Question>

What is the black couch facing?

<3D Scene>



training → 3D-Language alignment

point token

text token

Outputs:

<Answer>:

coffee table

1 Why Use 2D LVLM: Lack of Large-scale 3D-language Data

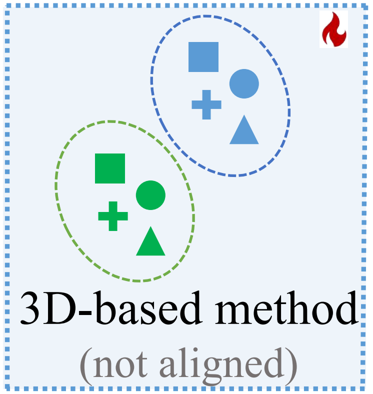
Strategies for achieving effective 3D-Language alignment

Inputs:

<Question>

What is the black couch facing?

<3D Scene>

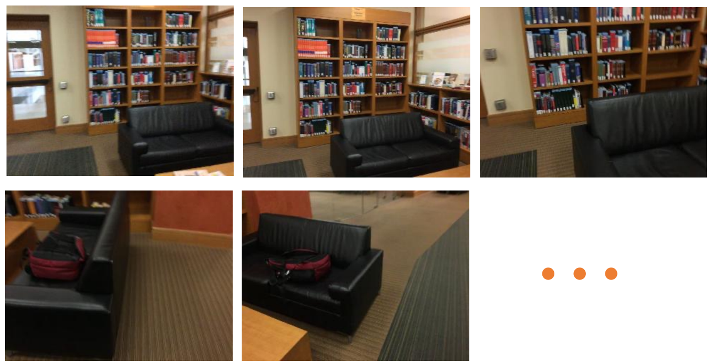


training → 3D-Language alignment

point token

text token

<2D Views>



Outputs:

<Answer>:

coffee table

1 Why Use 2D LVLM: Lack of Large-scale 3D-language Data

Strategies for achieving effective 3D-Language alignment

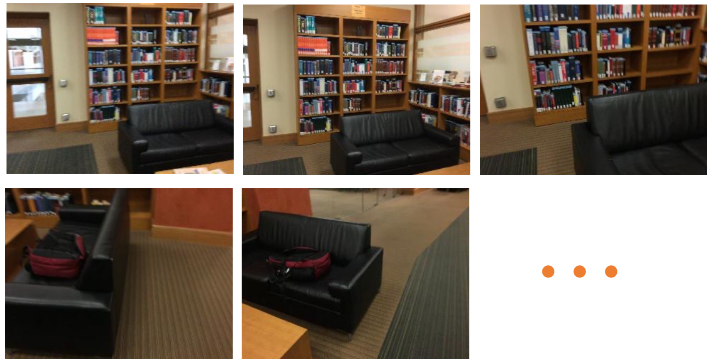
Inputs:

<Question>

What is the black couch facing?

<3D Scene>

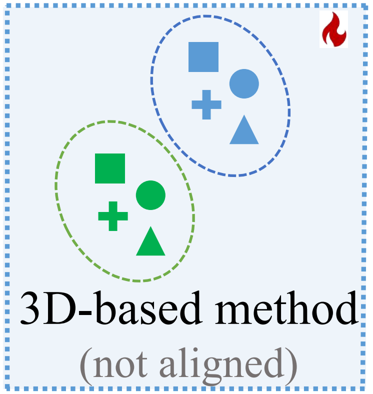
<2D Views>



Outputs:

<Answer>:

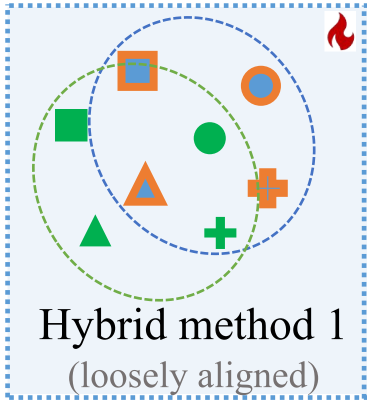
coffee table



training → 3D-language alignment

point token

text token



point token reconstructed from 2D views

image token

1 Why Use 2D LVLM: Lack of Large-scale 3D-language Data

Strategies for achieving effective 3D-Language alignment

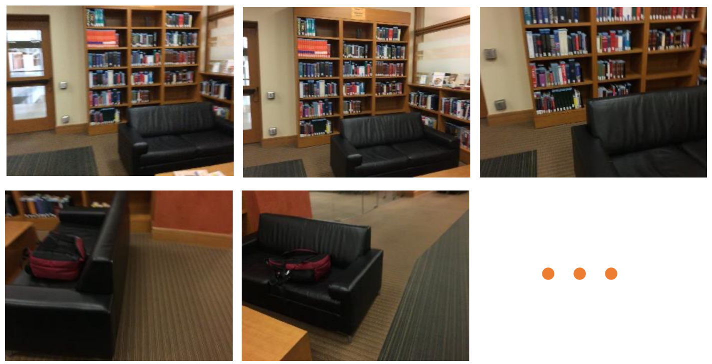
Inputs:

<Question>

What is the black couch facing?

<3D Scene>

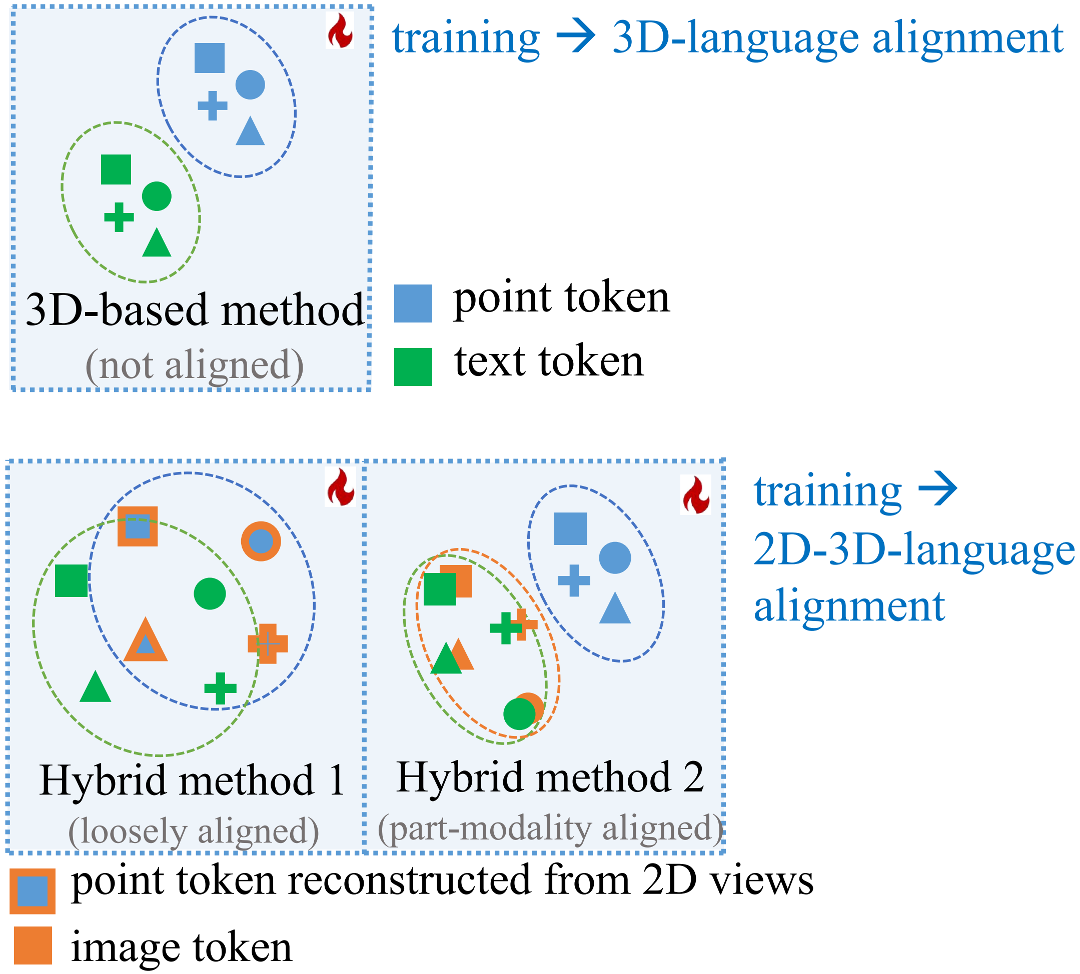
<2D Views>



Outputs:

<Answer>:

coffee table



1 Why Use 2D LVLM: Lack of Large-scale 3D-language Data

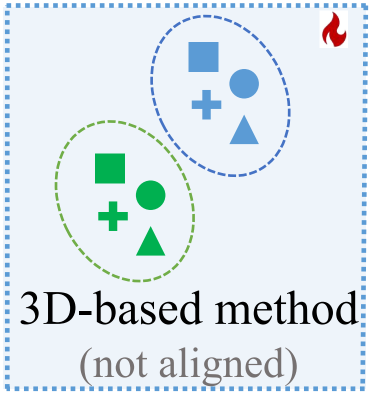
Strategies for achieving effective 3D-Language alignment

Inputs:

<Question>

What is the black couch facing?

<3D Scene>

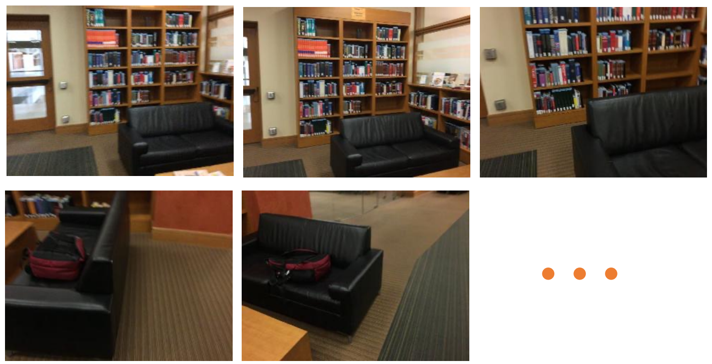


training → 3D-language alignment

point token

text token

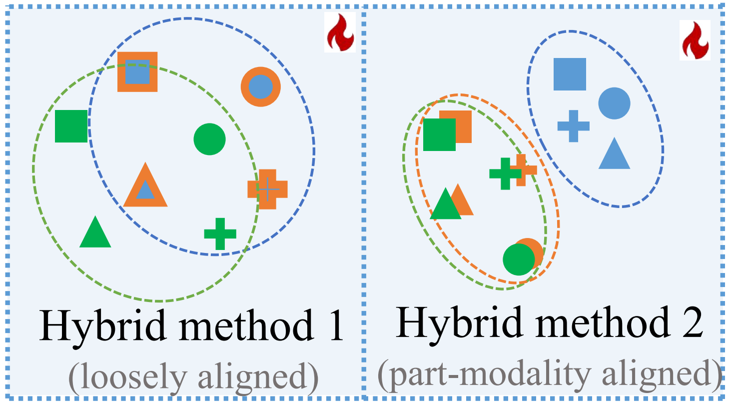
<2D Views>



Outputs:

<Answer>:

coffee table



training → 2D-3D-language alignment

point token reconstructed from 2D views

image token

1 Why Use 2D LVLM: Lack of Large-scale 3D-language Data

Strategies for achieving effective 3D-Language alignment

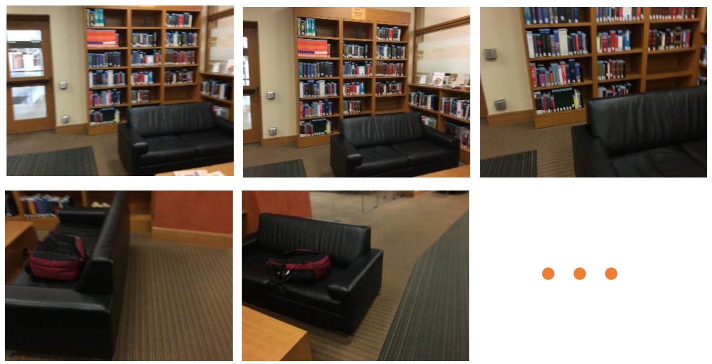
Inputs:

<Question>

What is the black couch facing?

<3D Scene>

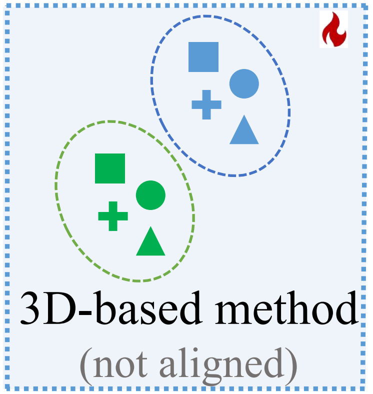
<2D Views>



Outputs:

<Answer>:

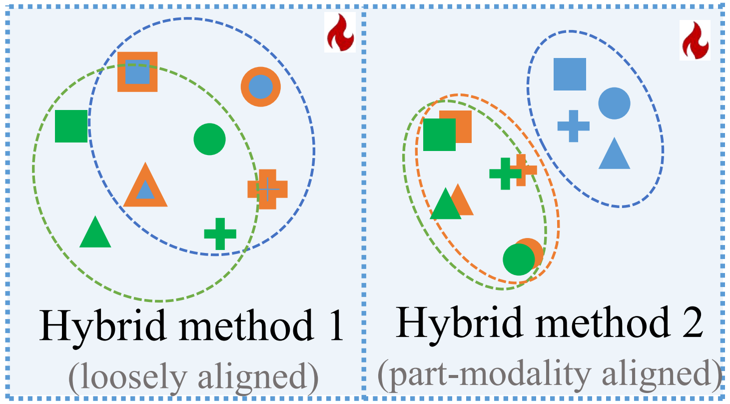
coffee table



training → 3D-language alignment

point token

text token



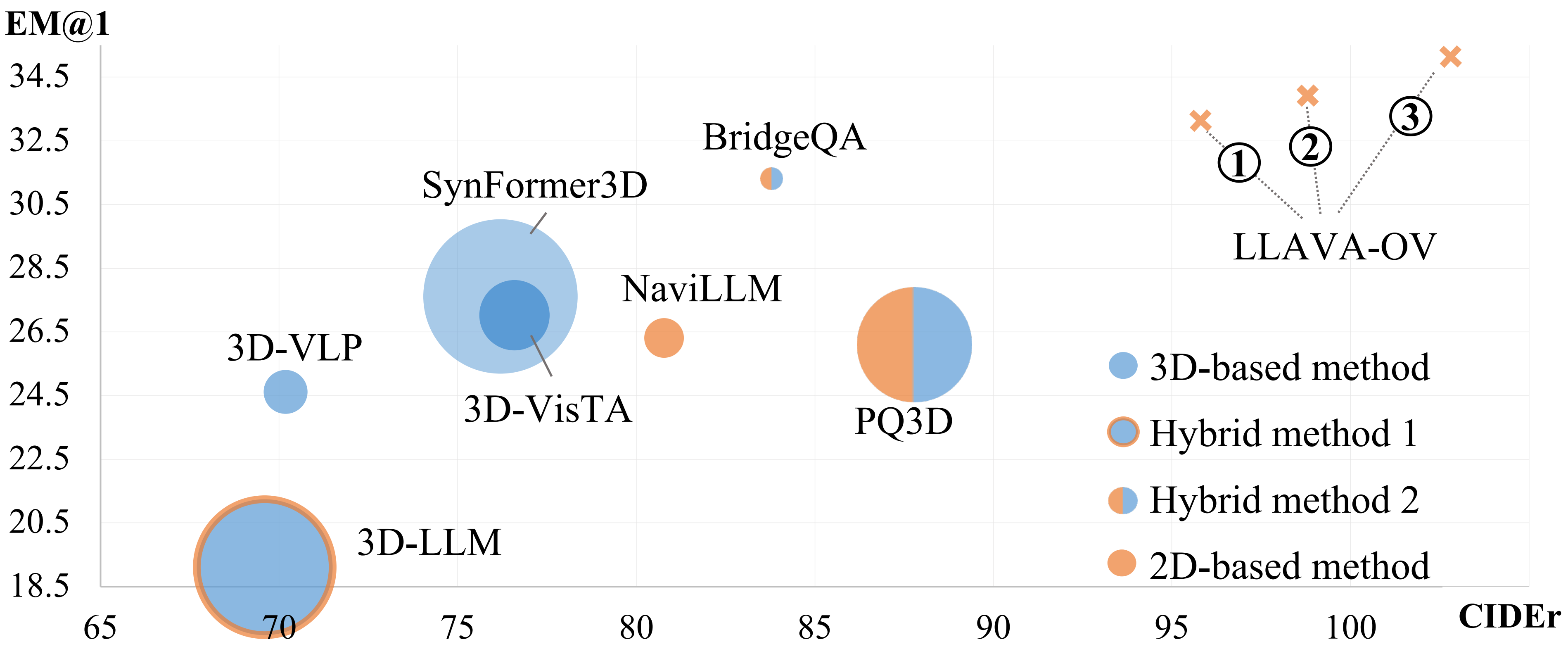
training → 2D-3D-language alignment

point token reconstructed from 2D views

image token

1 Why Use 2D LVLM: Lack of Large-scale 3D-language Data

3D-based and hybrid methods requires large amount of training data (large bubble size) but still leads to a poor 3D-QA performance:



Performance on the test set (with objects) of ScanQA

1 **Why Use 2D LVLM: Lack of Large-scale 3D-language Data**

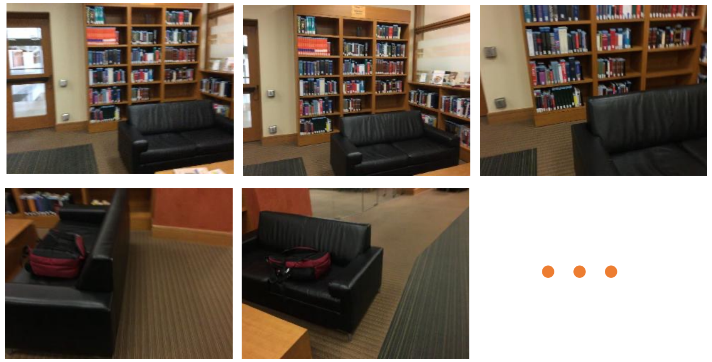
Using 2D LVLM in a zero-shot manner:

Inputs:

<Question>

What is the black couch facing?

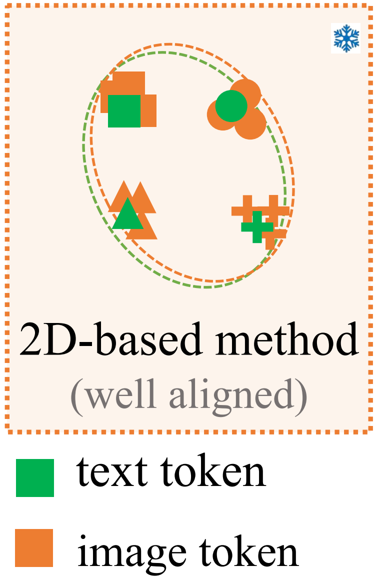
<2D Views>



Outputs:

<Answer>:

coffee table

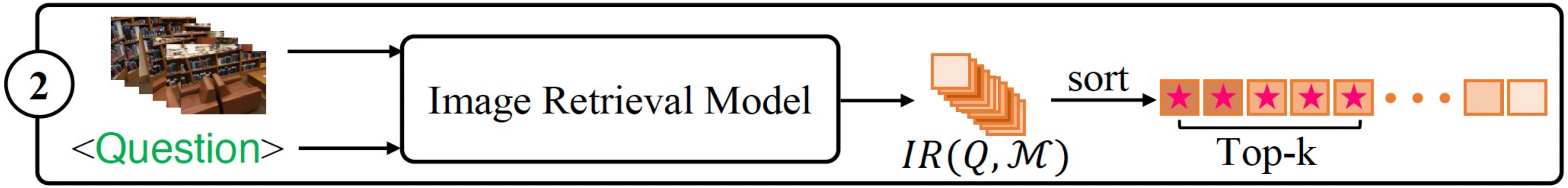


2 how to use 2D LVLM: zero-shot inference

Due to token limit, 2D LVLMs can only process a few views:

- uniform sampling:

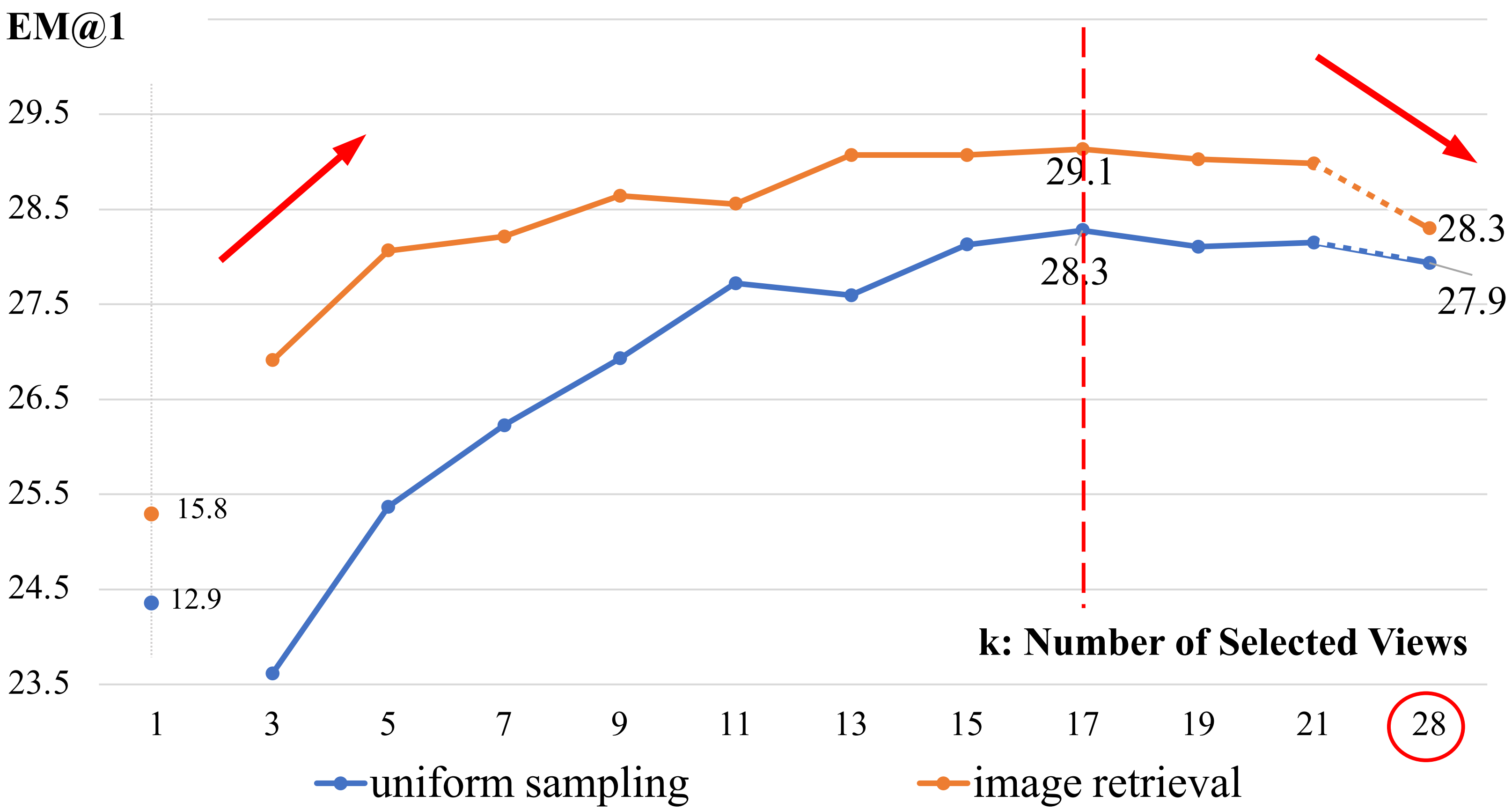
- image retrieval:



how to use 2D LVLM: zero-shot inference

add more views does not always help

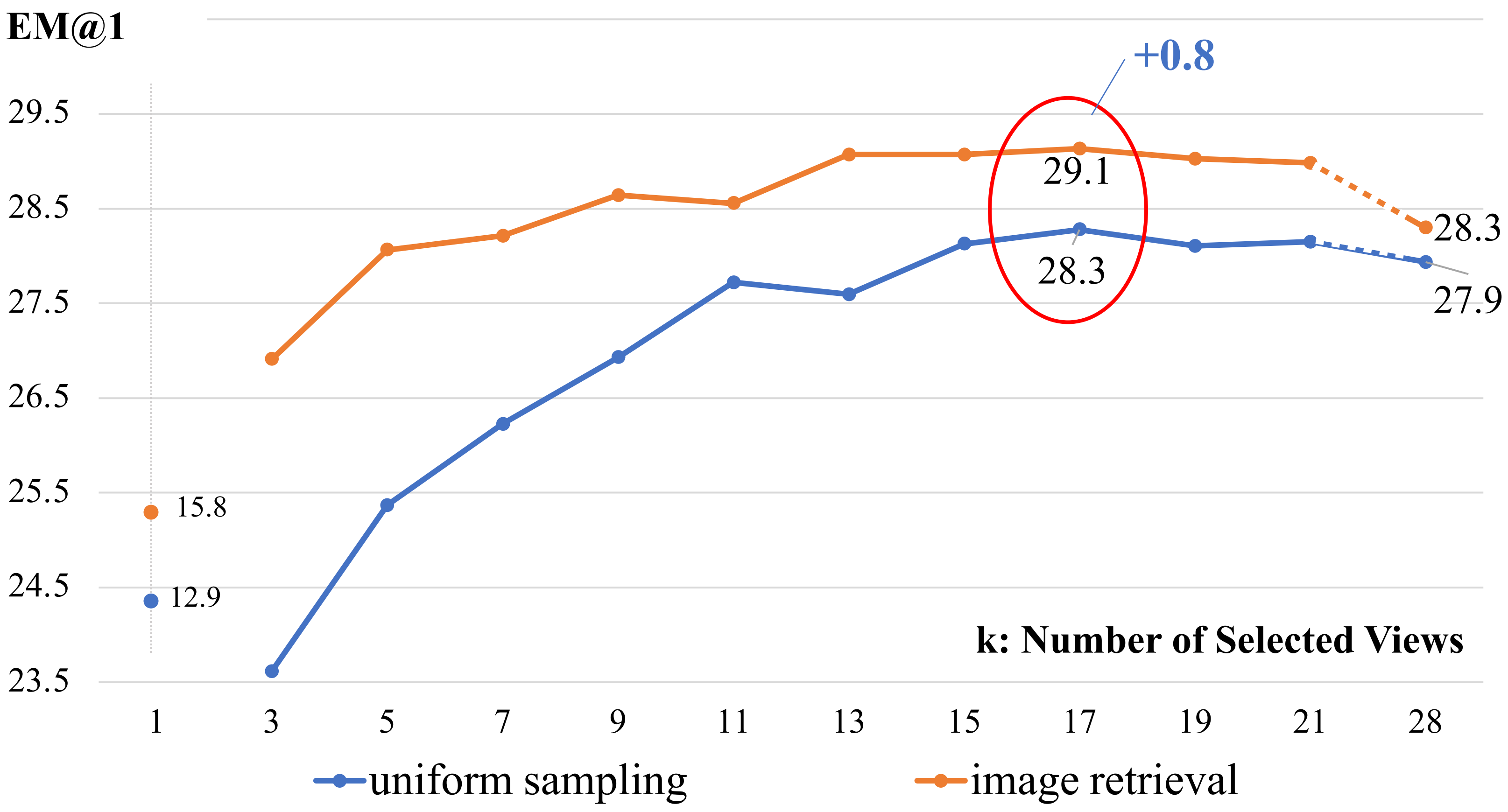
—in fact, it may degrade performance



ICML International Conference On Machine Learning

view selection is a key factor affecting performance

—— image retrieval vs. uniform sampling



2 how to use 2D LVLM: zero-shot inference

select critical and diverse views for 3D-QA

<Question>: What is the black couch facing?

<Answer>: coffee table

Uniform Sampling -- ignores question context

2 how to use 2D LVLM: zero-shot inference

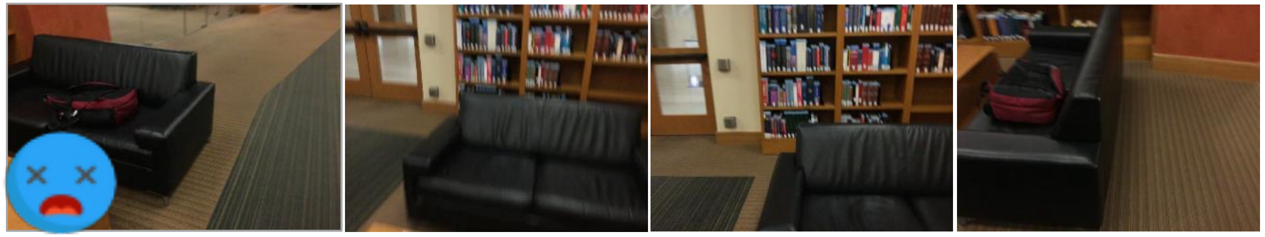
select critical and diverse views for 3D-QA

<Question>: What is the black couch facing?

<Answer>: coffee table

Uniform Sampling -- ignores question context

Image Retrieval – misses answer information



2 how to use 2D LVLM: zero-shot inference

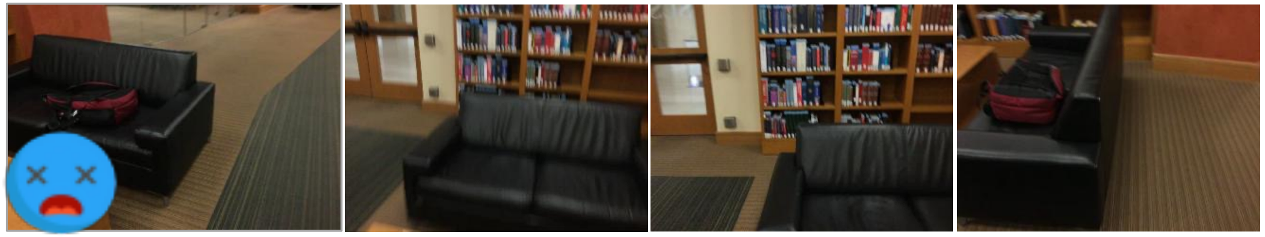
select critical and diverse views for 3D-QA

<Question>: What is the black couch facing?

<Answer>: coffee table

Uniform Sampling -- ignores question context

Image Retrieval – misses answer information



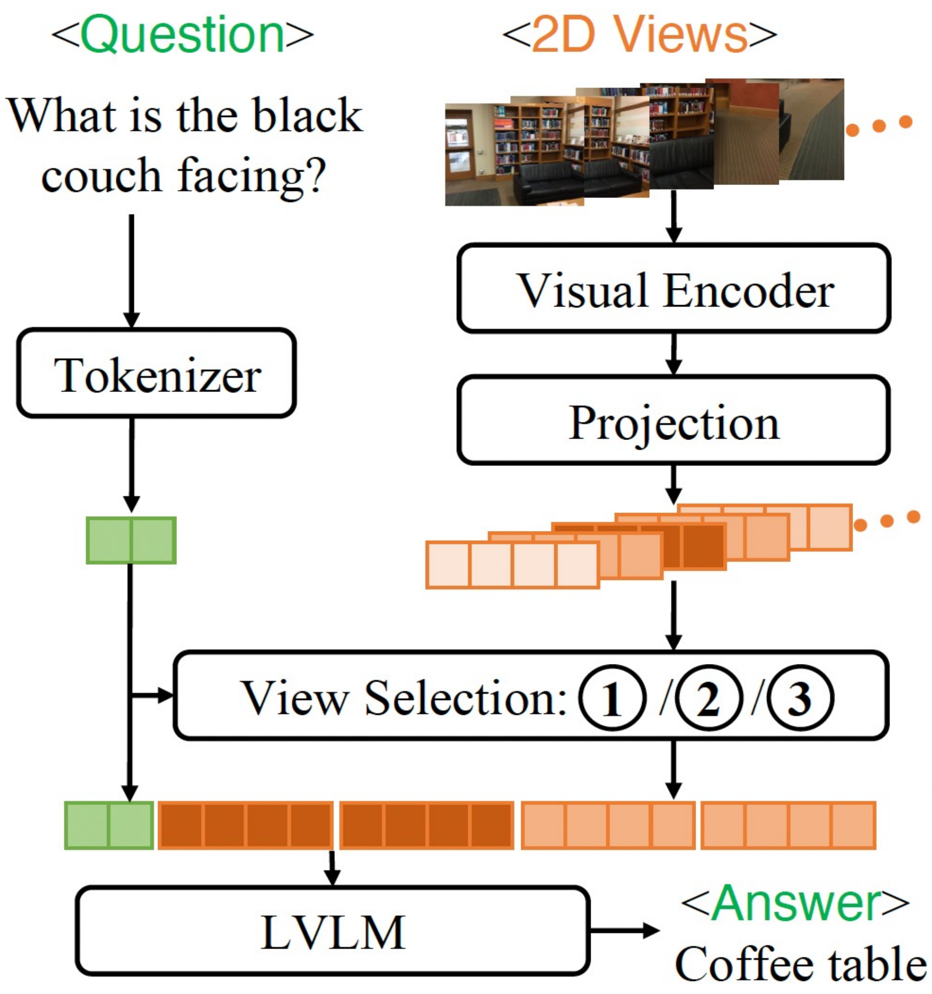
cdViews – "the black couch facing a coffee table" is included

Critical: contain information crucial for answering questions

Diverse: filter out the overlapping views

cdViews: Selecting critical and diverse Views

The pipeline of zero-shot 3D-QA:



option ①: uniform sampling

option ②: image retrieval

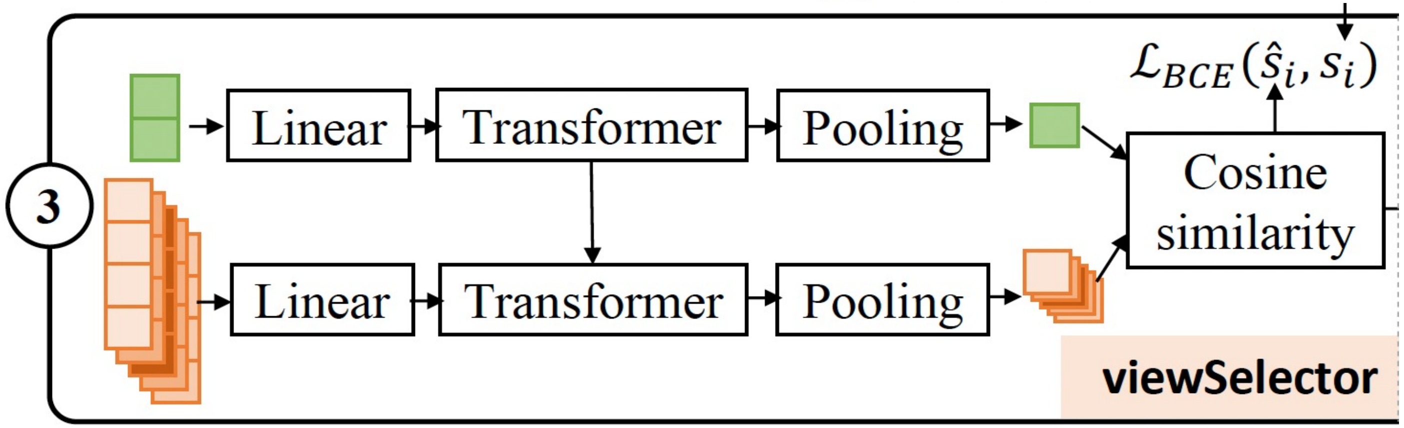
option ③: cdViews

3 cdViews: Selecting critical and diverse Views

Training Stage:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N'}\sum_{i=1}^{N'}\left(\hat{S}_i \log(S_i) + (1-\hat{S}_i)\log(1-S_i)\right)$$

viewAnnotator



positive views — label 1

negative views — label 0

uncertain views — disregard

viewSelector is trained in two steps:

1) Data annotation. viewAnnotator automatically label views as positive, negative, or uncertain.
2) Model training. viewSelector is then trained in a supervised manner using these labels.

3 cdViews: Selecting critical and diverse Views

The details of viewAnnotator:

Step 1: Caption Generation

<$Prompt_R$>: You are a helpful assistant. For each QA pair, generate a caption that describes the visual scene, fully incorporating relevant information from the question and answer.

<Question>: What is in the right corner of room by curtains? <Answer>: brown cabinet with tv sitting in it

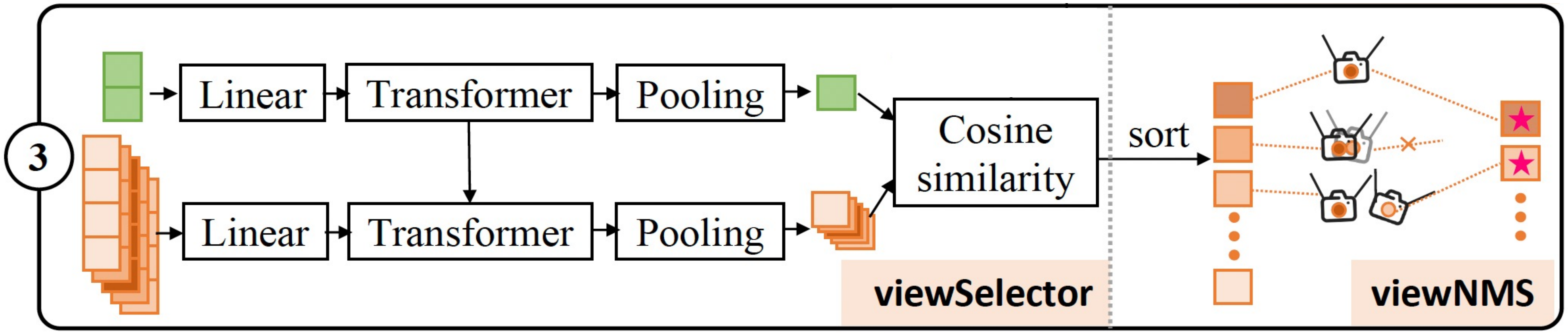
a brown cabinet with a television inside is located in the right corner of the room, near the curtains.

Step 2: View Matching

<$Prompt_M$>: You are given an image and a caption describing the visual content. Determine if the image matches the caption, and respond with one of the following options:
A. Yes, fully matches. B. No, does not match. C. Uncertain, insufficient or unclear information.

Positive | Negative

Inference Stage:

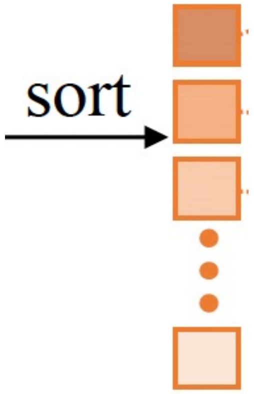


cdViews loads two modules for view selection:

1) viewSelector—prioritizes views most likely to contain answer-related information
2) viewNMS—removes redundant ones and improve the view diversity.

cdViews: Selecting critical and diverse Views

The details of viewNMS:

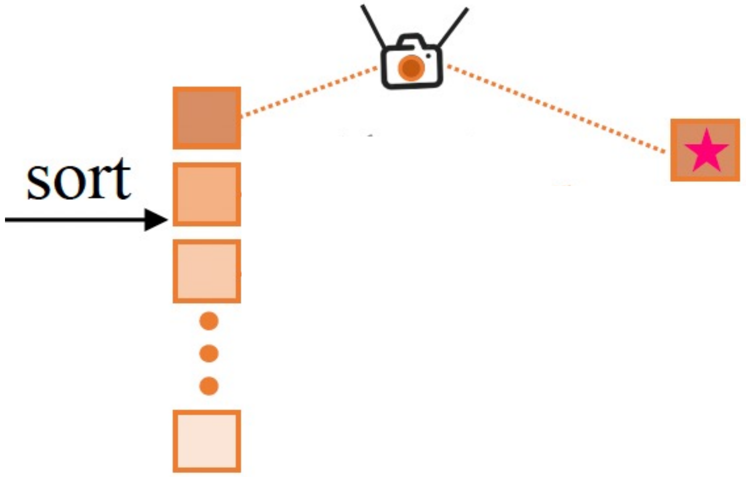


viewNMS operates in 3 steps:

1) Ranking views: sorts all views by their scores in descending order;

3 cdViews: Selecting critical and diverse Views

The details of viewNMS:

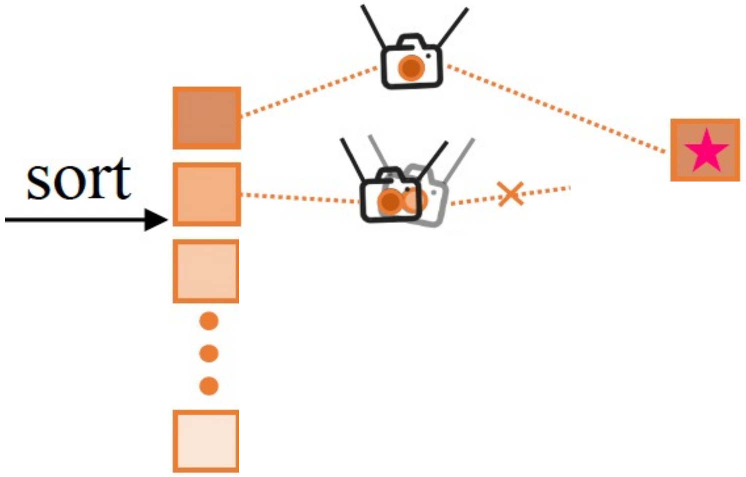


viewNMS operates in 3 steps:

1) Ranking views: sorts all views by their scores in descending order;
2) Initializing candidate views: selects the highest-scoring view as the initial set;

3 cdViews: Selecting critical and diverse Views

The details of viewNMS:

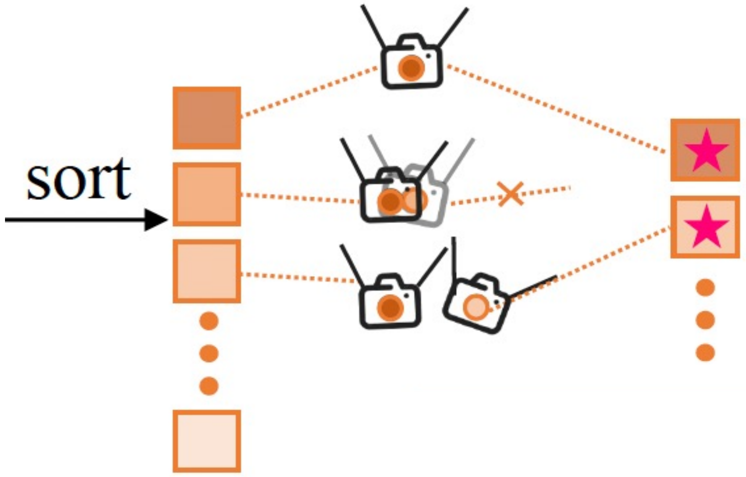


viewNMS operates in 3 steps:

1) Ranking views: sorts all views by their scores in descending order;
2) Initializing candidate views: selects the highest-scoring view as the initial set;
3) Adding diverse views sequentially: adding a view to the set if its distance from previously selected views exceeds a threshold T

3 cdViews: Selecting critical and diverse Views

The details of viewNMS:

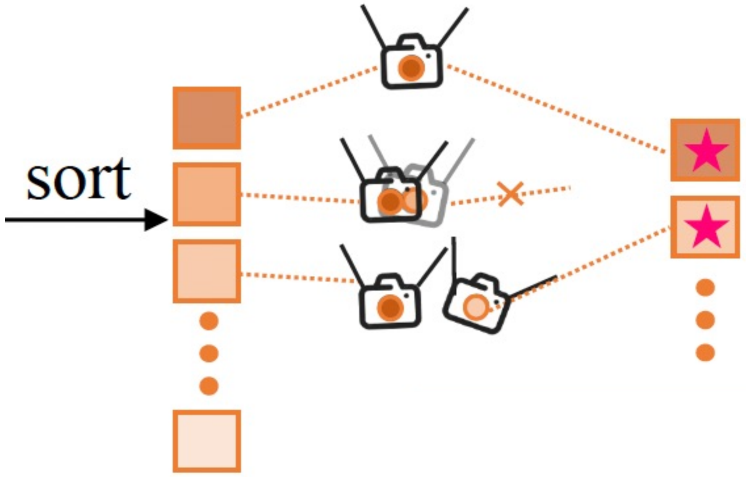


viewNMS operates in 3 steps:

1) Ranking views: sorts all views by their scores in descending order;
2) Initializing candidate views: selects the highest-scoring view as the initial set;
3) Adding diverse views sequentially: adding a view to the set if its distance from previously selected views exceeds a threshold T

3 cdViews: Selecting critical and diverse Views

The details of viewNMS:



View distance calculation (based on the camera parameters $[R|t]$):

1. orientation distance:

$$D_{ori}(V_i, V_j) = 2 \cdot \arccos(\|\mathbf{p}_i \cdot \mathbf{p}_j|)$$

where p_i is a quaternion representation of the orientation R_i

2. position distance:

$$D_{pos}(V_i, V_j) = ||\mathbf{t}_i - \mathbf{t}_j||$$

3. view distance:

$$D(V_i, V_j) = D_{pos}(V_i, V_j) + D_{ori}(V_i, V_j)$$

4 Experimental Results

Comparison with SOTA :

Method	Type	ScanQA EM@1	ScanQA BLEU-1	ScanQA ROUGE	ScanQA CIDEr	SQA EM@1
ScanQA (Azuma et al., 2022)	3D	23.5 / 20.9	31.6 / 30.7	34.3 / 31.1	67.3 / 60.2	45.3
SQA3D (Ma et al., 2022)	3D	-	-	-	-	47.2
3D-LLM (Hong et al., 2023)	3D	19.1 / -	38.3 / -	35.3 / -	69.6 / -	48.1
3D-VLP (Jin et al., 2023a)	3D	24.6 / 21.6	33.2 / 31.5	36.0 / 31.8	70.2 / 63.4	-
3D-VisTA (Zhu et al., 2023)	3D	27.0 / 23.0	-	38.6 / 32.8	76.6 / 62.6	48.5
SIG3D (Man et al., 2024a)	3D	-	-	-	-	52.6
SynFormer3D (Yang et al., 2024)	3D	27.6 / 24.1	-	39.2 / 33.3	76.2 / 62.7	-
LL3DA (Chen et al., 2024a)	3D+2D	-	-	38.2 / 35.2	78.2 / 70.3	-
PQ3D (Zhu et al., 2025)	3D+2D	26.1 / 20.0	43.0 / 36.1	-	87.8 / 65.2	47.1
BridgeQA (Mo & Liu, 2024)	3D+2D	31.3 / 30.8	34.5 / 34.4	43.3 / 41.2	83.8 / 79.3	52.9
LLAVA-OV + $\mathcal{F}_{\text{uniform}}$	2D	33.1 / 33.5	43.2 / 44.2	46.9 / 46.6	95.8 / 93.3	53.5
LLAVA-OV + $\mathcal{F}_{\text{retrieval}}$	2D	33.9 / 34.6	44.8 / 46.1	48.3 / 48.7	98.8 / 97.7	55.0
LLAVA-OV + $\mathcal{F}_{\texttt{cdViews}}$	2D	**35.1 / 35.6**	**46.1 / 47.2**	**49.7 / 49.5**	**102.8 / 100.4**	**56.9**
margin over the compared best	-	3.8 ↑ / 4.8 ↑	3.1 ↑ / 9.1 ↑	6.4 ↑ / 8.3 ↑	15.0 ↑ / 21.1 ↑	3.9 ↑

4 Experimental Results

Ablation studies — cdViews components

LLAVA-OV	viewSelector	viewNMS	Best EM@1	Optimal number of views k
+ $\mathcal{F}_{\text{uniform}}$	-	-	28.3	17
+ $\mathcal{F}_{retrieval}$	-	-	29.1	17
+ $\mathcal{F}_{\texttt{cdViews}}$	✓	-	29.7	17

viewSelector: improves by 1.4% over the uniform sampling baseline, which validate that it can effectively prioritizes critical views.

Ablation studies — cdViews components

LLAVA-OV	`view` `Selector`	`view` `NMS`	Best EM@1	Optimal number of views k
+ $\mathcal{F}_{\text{uniform}}$	-	-	28.3	17
+ $\mathcal{F}_{retrieval}$	-	-	29.1	17
+ $\mathcal{F}_{retrieval}$	-	✓	29.2	9
+ $\mathcal{F}_{\texttt{cdViews}}$	✓	-	29.7	17
+ $\mathcal{F}_{\texttt{cdViews}}$	✓	✓	30.1	9

viewNMS: reduces the input to just 9 views—almost half the visual token length—without reducing the performance but further boosting EM@1 by 0.4%.

Conclusion

Research Problems:

- Lack of large-scale 3D-language dataset in 3D-QA
- how to effectively use 2D LVLM for 3D-QA in a zero-shot manner

Contributions:

- explore the use of 2D LVLM to address 3D-QA in a zero-shot manner
- introduce cdViews to capture critical and diverse views
- achieves state-of-the-art performance on two 3D-QA benchmarks

ICML
International Conference
On Machine Learning

Thank you

Paper: https://arxiv.org/pdf/2505.22143

Code: https://github.com/fereenwong/cdViews