VideoNavQA: Bridging the Gap between Visual and Embodied Question Answering

Cătălina Cangea[1] Eugene Belilovsky[2] Pietro Liò[1] Aaron Courville[2,3]

[1]*University of Cambridge* [2]*Mila, Université de Montréal* [3]*CIFAR Fellow*

Why is a shift in EQA perspective useful?

Embodied Question Answering requires an agent in a rich 3D environment to act based solely on egocentric input to answer a question.

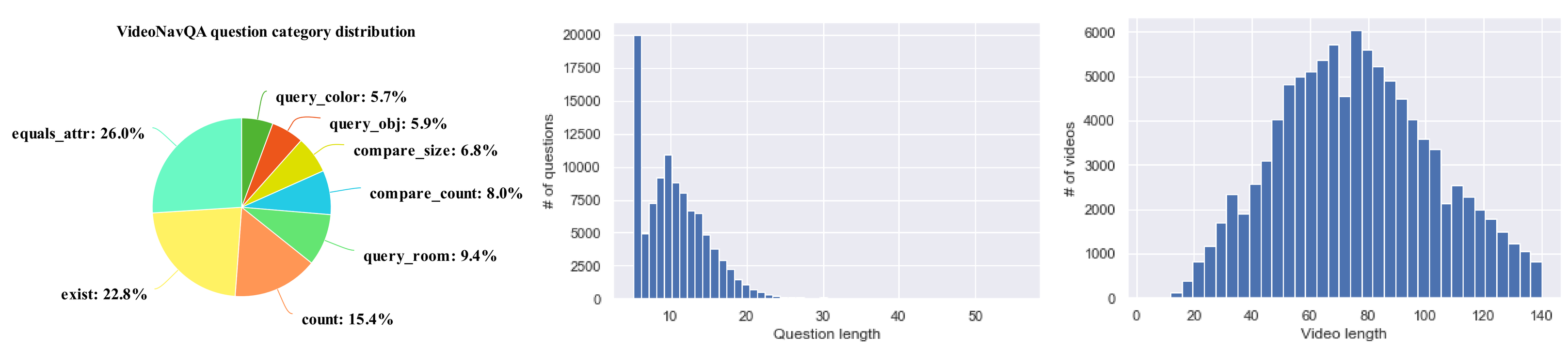
Learning to combine scene understanding, navigation and language understanding is needed to perform complex reasoning, and **initial advancements have shown EQA might be too challenging for existing imitation learning and reinforcement learning approaches**.

We construct VideoNavQA to investigate EQA-style task feasibility:

- assessing QA performance from **nearly-ideal navigation paths**
- considering **much more complex and varied questions**:

EQA-v1 (Q types: 4)	What room is the *<OBJ>* located in? What color is the *<OBJ>* in the *<ROOM>*?
VideoNavQA (Q types: 28)	Are both *<attr1><OBJ1>* and *<attr2><OBJ2>* *<color>*? How many *<attr>* *<OBJ>* are in the *<ROOM>*? Is there *<art>* *<attr>* *<OBJ>*?

Dataset statistics



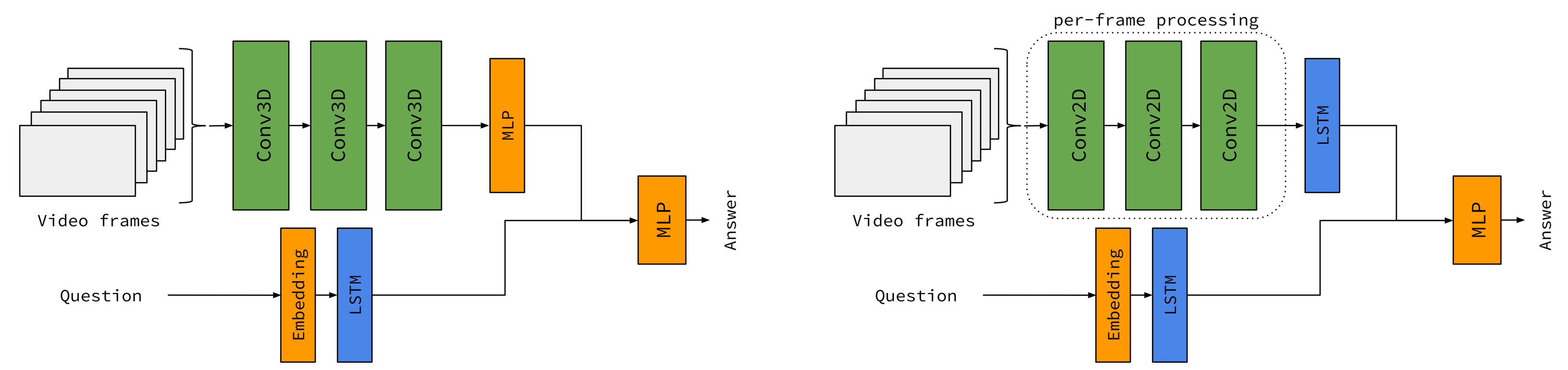
Left: Proportions per question category.
Middle: Question lengths (max = 56). *Right:* Video lengths (max = 140).

8 question categories, 28 question types, 70 possible answers.

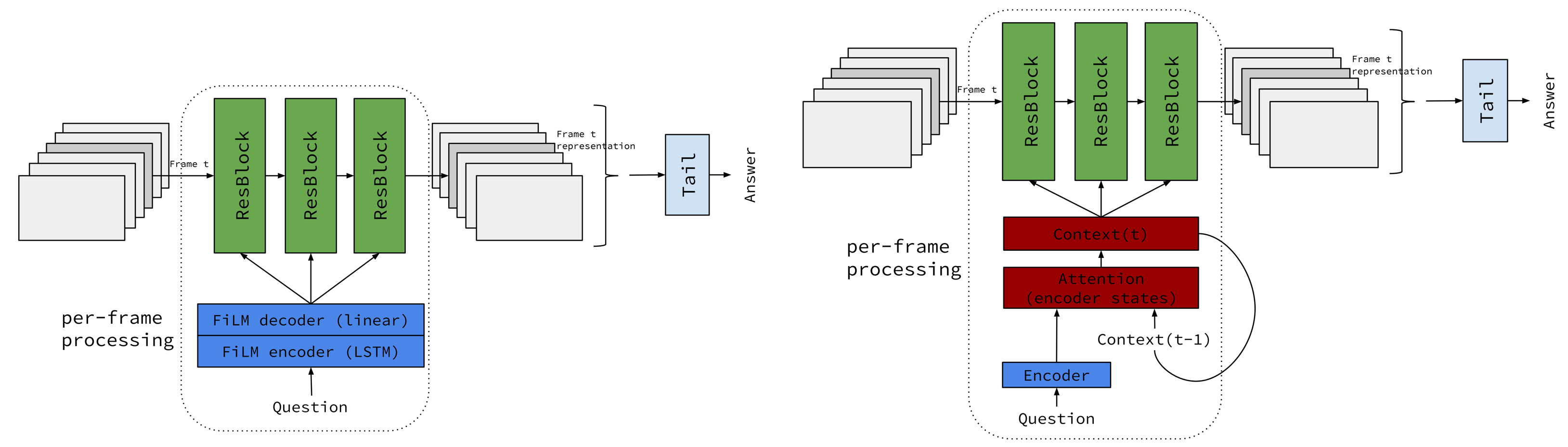
	# of houses	# of samples
Train	620	84807
Validation	65	8734
Test	55	7430

Generalized VQA models

Our benchmark reimagines the EQA task while requiring **a smaller degree of fusion among different classes of methods**. The architectures used to obtain initial results are several essential baselines and **new models inspired by previous successes** in VQA and computer vision.



Left: **Concat-CNN3D** processes the entire video. *Right:* **Concat-CNN2D** aggregates frame features via an LSTM. Both merge the result with the question embedding.



Left: **Per-frame FiLM**. Video frames are processed separately by ResBlocks, then all features are aggregated by the classifier to answer the question. *Right:* **Temporal multi-hop**. Each video frame is processed by the ResBlocks: FiLM parameters are computed from the current attention context, which is initialized with the one from the previous frame. Temporal summarization is achieved via global max-pooling.

We extend Compositional Attention Networks (**MAC**) by applying a 2D-CNN to each video frame and feeding the resulting representation at each time step to a MAC model—this performs iterative inference with attention over the frame. Results are integrated over time via an LSTM.

Are we actually using the visual input?

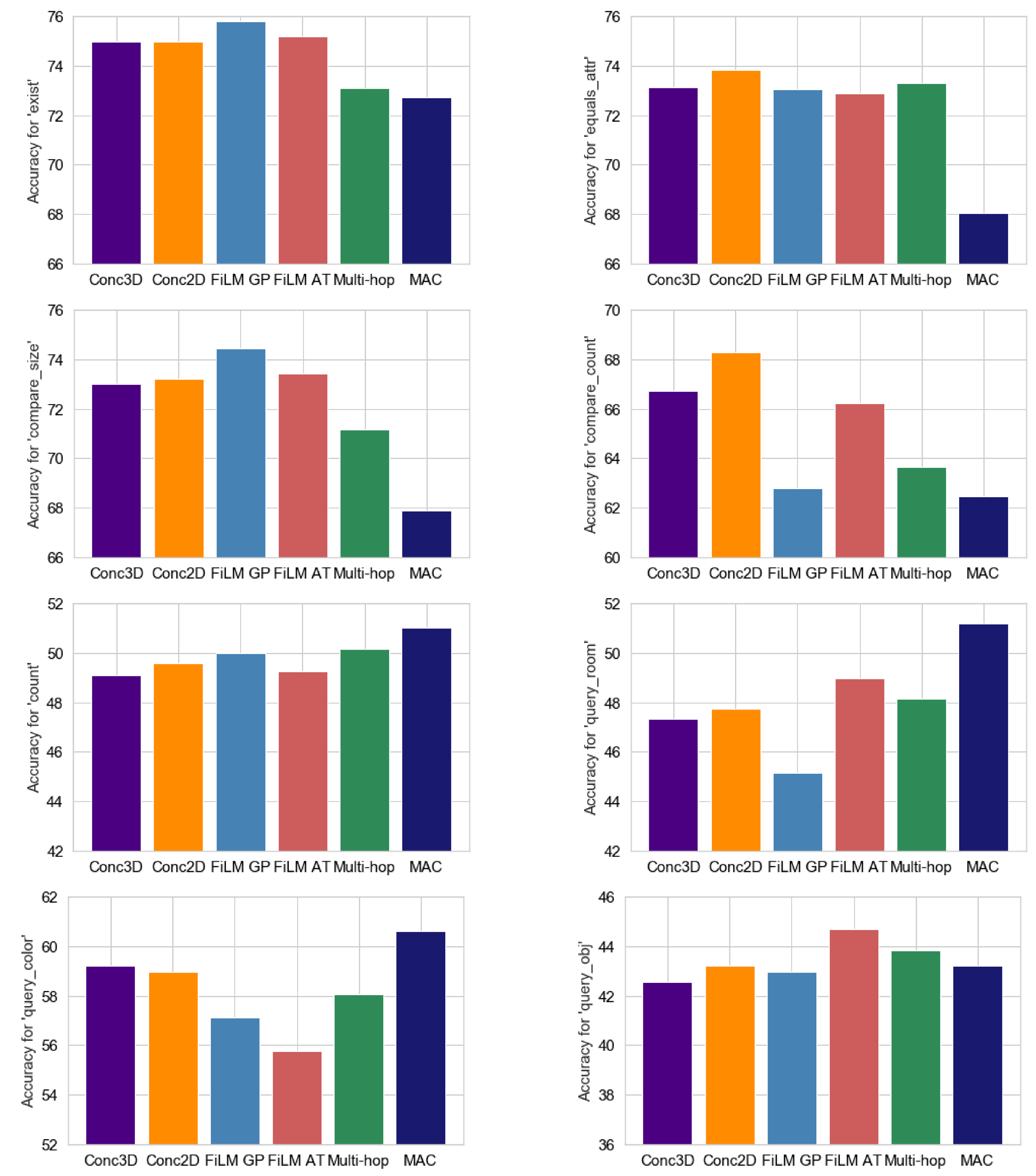
Question-only baselines have been surprisingly effective in EQA, often performing better than complex approaches. We evaluate two simple yet powerful models: a 1-layer **LSTM** and a bag-of-words (**BoW**):

- reveal **inherent biases** in the environment distribution
- **performance lower bound** for models that exploit visual information

Overall performance

Model	All	Yes/No	Other	Num
BoW	49.02	57.67	30.57	40.21
LSTM	56.49	68.36	35.27	38.90
Concat-CNN3D	64.00	72.99	49.12	49.10
Concat-CNN2D	64.47	73.50	49.20	49.59
FiLM-GP	63.79	72.91	47.71	50.00
FiLM-AT	64.08	72.93	49.54	49.26
Temporal multi-hop	63.53	71.81	49.54	50.16
MAC	62.32	69.02	51.37	50.99

Detailed analysis per question category



https://arxiv.org/abs/1908.04950