Do Visual Imaginations Improve Vision-and-Language Navigation Agents?

Supplementary Material

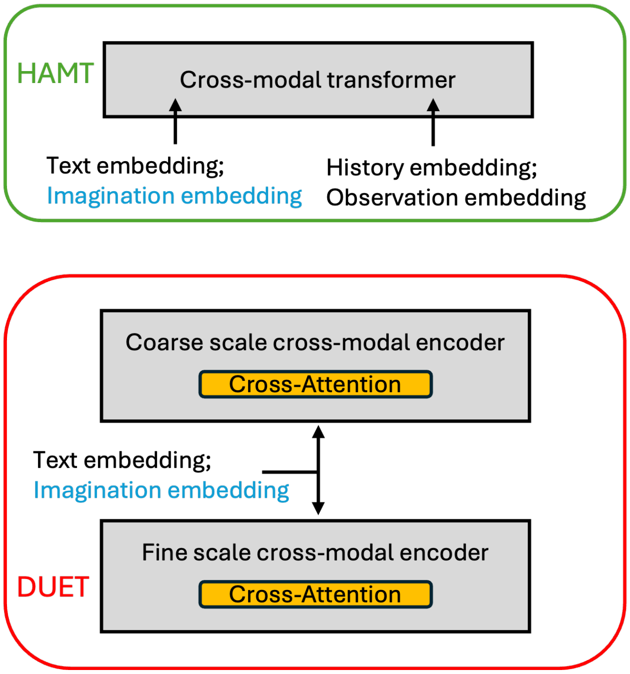


Figure 5. Integration of imagination embeddings to HAMT and DUET. The imagination embeddings are concatenated with language embeddings before passing through cross-modal encoders.

6. Model details

6.1. Integration with base agents

We show our integration mechanism in detail for HAMT and DUET in Fig. 5. In HAMT, the imagination embedding vector h is concatenated with language embeddings and passed to the language modality branch of HAMT's CMT. The visual modality branch with observations and history along with the action prediction network is retained as is. In DUET, the imagination embedding vector h is concatenated with language embeddings and passed to the coarse-scale cross-modal encoder to perform global action prediction, and to the fine-scale cross-modal encoder to perform local action prediction. The fusion between the dual-scale action predictions along with rest of the architecture is retained as is.

6.2. Training routine

In order to mitigate catastrophic forgetting, we train our agent for 100k iterations in three stages:

- First, we train only the imagination encoder MLP along with type embeddings t^{Im} at a learning rate of $1e-4$ for 25% of iterations. Rest of the modules from the base agent are frozen.
- Then, all the modules are trained jointly for the next 25% of iterations. The imagination encoder parameters are trained at a learning rate of $5e-5$ and the base agent parameters are trained at $1e-6$.
- Finally, all the parameters are trained at a common learning rate of $1e-6$ for rest of the training.

6.3. MLP architecture

Our MLP consists of three fully connected (FC) layers with ReLU activations after the first two layers. Our input dimension is 768, hidden dimension is 512 and output dimension is 768. We apply a dropout layer with rate 0.15 to the input and omit bias terms in all layers.

7. Imagination guidance prompts

We use positive prompts to guide the generations towards indoor environments and negative prompts to exclude concepts such as humans and outdoor environments. The complete prompts are listed below:

- Positive prompts: indoor, house, realistic, real estate.
- Negative prompts: outdoor, text, humans, man, woman, boy, girl, collage.

8. Additional experiments

Table 8. Visual *vs.* textual representations; Training with mean sub-instruction embeddings in place of imagination embeddings leads to a performance drop albeit better than baseline implying the advantages of visual imaginations.

	SR↑	SPL↑
Baseline (HAMT)	66.24	61.51
HAMT-Imagine (ours)	67.26	62.02
Sub-instrs only	66.67	61.50

Visual representations of landmarks outperform textual representations. We study the effect of replacing visual imaginations with textual sub-instructions in this experiment. To do this, we train a HAMT based agent similar to HAMT-Imagine but in place of imagination embeddings, we use mean sub-instruction embeddings. In Tab. 8, we notice textual representations while leading to better navigation performance than baseline is still inferior to visual imaginations by 0.59 SR and 0.52 SPL. This might imply imaginations play a complementary role to language in our

Table 9. Design ablations for imagination encoder. We observe best performance when imaginations are encoded by an MLP and concatenated along with instruction encodings.

Design	SR↑	SPL↑
HAMT-Imagine	67.26	62.02
Transformer	66.75	61.64
Visual concat	66.54	61.38

Table 10. Early fusion *vs*. late fusion. Our early fusion leads to better navigation success.

Design	SR↑	SPL↑
DUET-Imagine (early fusion)	72.12	60.48
LAD late fusion	71.73	60.44

setting.

Simple MLP is a sufficient imagination encoder when concatenated with language. We experiment with alternate designs for encoding visual imaginations in Tab. 9. We use a transformer (row 2) with positional encodings to encode imaginations before passing them to the cross-modal encoder in HAMT. MLP as imagination encoders leads to a better performance (rows 1, 2) in our setting. One possible explanation is that having fewer parameters can help reduce overfitting, which VLN agents are susceptible to due to limited data availability. Additionally, imaginations when concatenated with instructions as opposed to visual observations improves effectiveness of our agent (rows 1, 3). We hypothesize that imaginations act as "visual instructions" such that concatenating them with instruction might aid in relevant inductive biases. Finally, we compare our early fusion approach of integrating imagination features with that of LAD's [26] late fusion using our DUET-imagine agent. LAD fuses imagination features with DUET's global features using a node-specific learned weight to compute an action distribution. We incorporate their approach to our DUET-imagine agent and contrast it with our early fusion. We report in Tab. 10, early fusion provides an improvement of 0.39 SR and 0.04 SPL.

9. Qualitative visualizations

First, we show additional qualitative visualizations of top attended concepts in Fig. 6. In the first example (row 1), the imagination captures noun phrases "black table" and "chair". The imagination strongly attends to related language tokens "black", "table" and a noun phrase from a different sub-instruction "out". The top attended observation images by the selected attention head contains black table and chairs. As can be partially seen in the observation images, the neighborhood contains other black objects that can mislead an agent hinting at the potential of imaginations in disambiguating similar concepts. In the second example, the imagination captures a stove and kitchen. The top attended language tokens by a selected attention head are kitchen and stove. Its corresponding top attended observation tokens capture the same concepts as well.

We also illustrate a sample trajectory of our HAMT-Imagine agent in comparison with the baseline HAMT agent in Fig. 7. We consider sample 431_1 from R2R val-unseen with instruction "Walk into the bedroom area. Walk passed the bed and through the door. Walk down the hallway and into the bedroom with the striped bed backboard and golden blanket laying on top." The baseline HAMT agent stops prematurely adjacent to a different bedroom along the path. Our agent, which is provided a synthesized imagination containing a striped backboard and golden blanket, successfully continues past the incorrect bedroom and stops at the correct bedroom. We hypothesize our agent is able to use the imagination to disambiguate between similar concepts in this example.

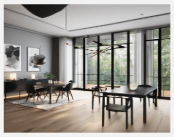
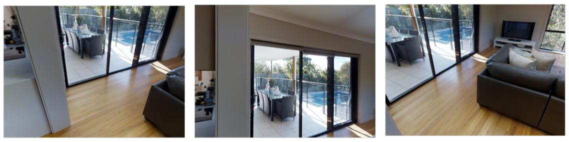
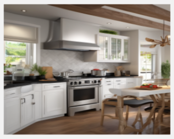
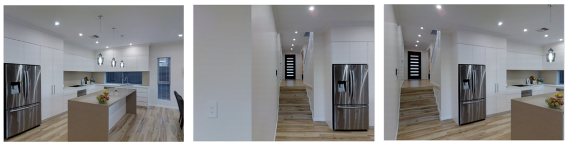
Sub-instruction	Imagination	Top Attended Tokens	Top Attended Observations
And wait near the black table and chair		black, out, table	
Stop nest to the stove and island in the kitchen		kitchen, stove, .	

Figure 6. Additional qualitative examples illustrating the role of imaginations as pivots between language and observation images. The first example (row 1) illustrates the potential application of imaginations in disambiguating destinations with similar looking objects (black colored objects). The second example (row 2) showcases the potential ability of imaginations to act as a pivot between language and observations of kitchen and stove.

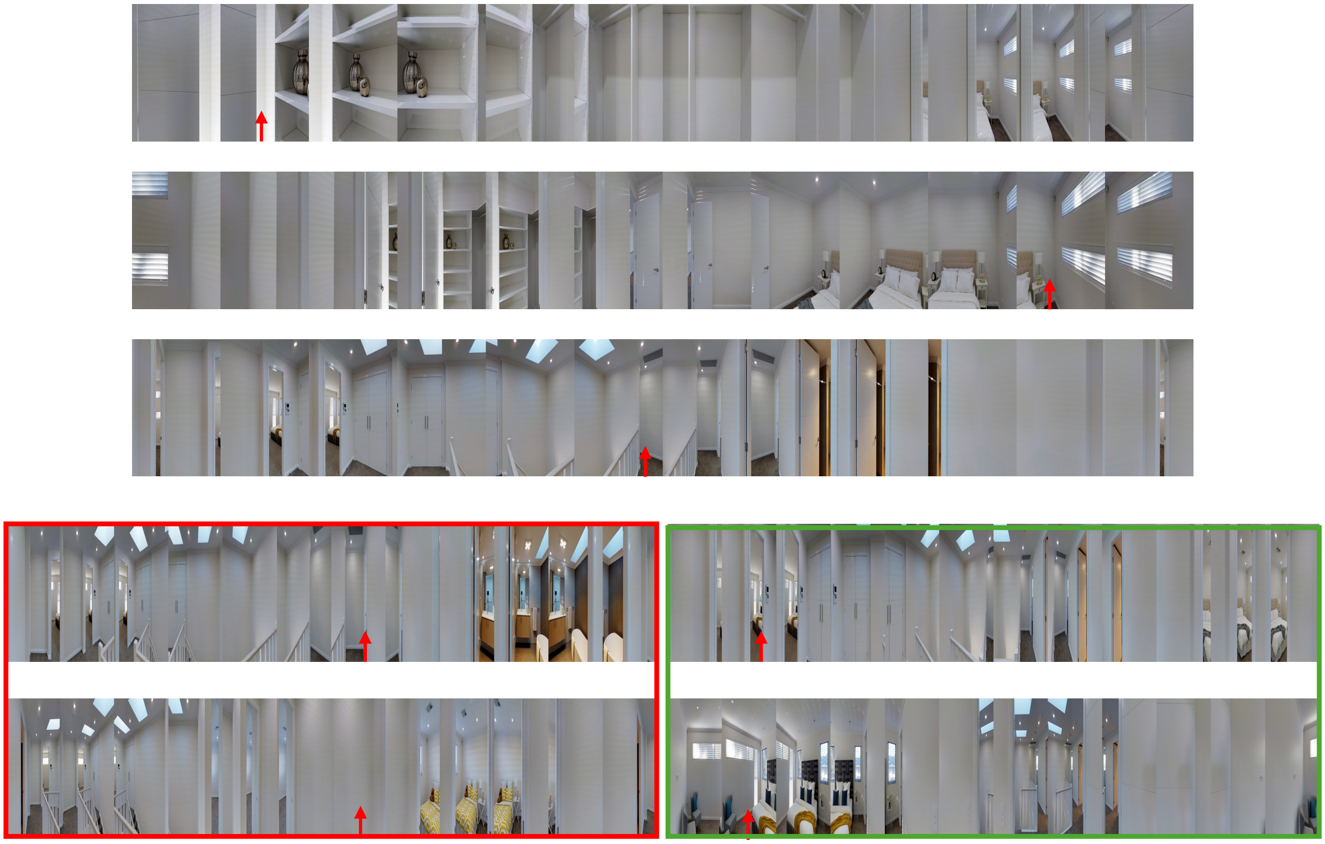


Figure 7. Qualitative visualization of trajectory from HAMT and HAMT-Imagine for sample 431_1. In the first three timesteps, both the agents are aligned. However, HAMT (left) is unable to decide between the two bedrooms in close vicinity whereas our agent HAMT-Imagine (right) is able to disambiguate between similar concepts to arrive at the correct bedroom with a golden blanket.