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

Akhil Perincherry[1], Jacob Krantz, Stefan Lee[1]
[1]Oregon State University
perincha@oregonstate.edu, jkrantz989@gmail.com, leestef@oregonstate.edu

Abstract

Vision-and-Language Navigation (VLN) agents are tasked with navigating an unseen environment using natural language instructions. In this work, we study if visual representations of sub-goals implied by the instructions can serve as navigational cues and lead to increased navigation performance. To synthesize these visual representations or "imaginations", we leverage a text-to-image diffusion model on landmark references contained in segmented instructions. These imaginations are provided to VLN agents as an added modality to act as landmark cues and an auxiliary loss is added to explicitly encourage relating these with their corresponding referring expressions. Our findings reveal an increase in success rate (SR) of ∼1 point and up to ∼0.5 points in success scaled by inverse path length (SPL) across agents. These results suggest that the proposed approach reinforces visual understanding compared to relying on language instructions alone. Code and data for our work can be found at https://www.akhilperincherry.com/VLN-Imagine-website/.

1. Introduction

In this work, we examine whether providing visual imagery corresponding to described landmarks improves the performance of agents following natural language navigation instructions. Consider the natural language navigation instruction presented in Figure 1 which asks an agent to "*Go straight, take a left at the pool table to enter the kitchen. Walk to the bedroom and stop*". This instruction provides unconditional action directives like "*go straight*" but also frequently conditions the given directions on visual landmarks in the scene such as the pool table, kitchen, and bedroom. Standard approaches to vision-and-language navigation tasks rely on learned cross-modal alignment mechanisms to *implicitly* associate these noun phrases with their visual referents during navigation. However, text-to-image generation models have improved to the point that producing imagery matching the semantics of these visual references prior to navigation is plausible. We show three such

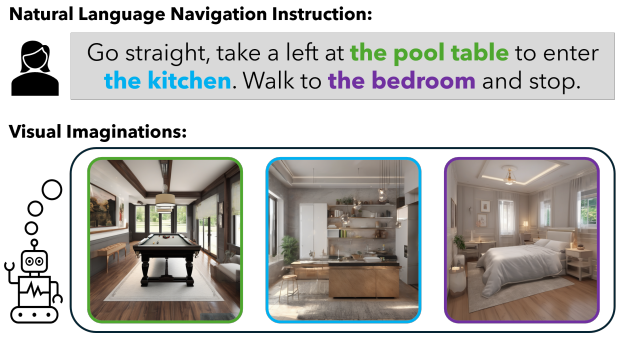


Figure 1. Illustration of visual imaginations. (Top) A natural language instruction specifying sub-goals pool table, kitchen, and bedroom. (Bottom) Visual imaginations of landmarks pool table, kitchen and bedroom referenced by the sub-goals in the instruction. In our work, we study if these visual imaginations generated using text-to-image models can improve performance in VLN.

generations for our running example in Figure 1. Drawing an analogy to the substantial work in cognitive science on the impact of mental imagery on task performance, we refer to these generated images as visual *imaginations* and study whether providing them in addition to corresponding language-based instructions can improve the performance of vision-and-language navigation agents.

But why might we believe visual imaginations could be beneficial? In short – text-to-image models have broad knowledge, and learning to perform semantic matching in the image domain may be an easier task than language grounding. Associating noun phrases and their visual referents is a sub-task in vision-and-language navigation (VLN) that most works address through extensive pretraining of model components on web-scale image caption datasets – e.g., by initializing image and language encoders with CLIP [34] models. While this provides a useful warm-start, it is unclear how well the remaining cross-modal components that drive navigation retain these capabilities after downstream VLN training. Downstream datasets may cover only a small portion of relevant objects and visual grounding remains a weakness in many existing VLN models [43, 47].

Directly synthesizing visual imaginations using off-the-shelf text-to-image models offers alternative training and inference mechanisms. Synthesized images have known correspondences to noun phrases in instructions that can be leveraged to reinforce visual grounding during downstream task training through auxiliary losses. Further, matching a visual imagination with environment observations reduces to an image-to-image matching task – offering the potential to use visual imaginations as a convenient pivot between language and vision. This may prove especially effective for out-of-distribution landmarks that have limited or no support during downstream task training. For example, novel instructions referencing "butterfly sculptures" or "Pulp Fiction posters" are unlikely to be well-grounded in the training set but produce semantically relevant visual imaginations that could then be matched to observations.

Though we do not imply any direct connections, our work is also inspired by cognitive science studies which indicate mental imagery can serve as valuable cues in addition to linguistic stimuli. Several works supporting dual-coding theory [29–31] find that concrete imagery tends to be more effective for certain types of concept learning compared to linguistic stimuli alone. Further, findings from human and animal studies [23, 36, 38, 39] suggest that imagining snapshots of the environment or potential trajectories before navigating can improve decision making.

To investigate our key question, we augment agents with visual imaginations for the popular Room-2-Room (R2R) [4] and REVERIE [33] tasks. We propose a VLN agent agnostic procedure to incorporate visual imaginations as additional inputs and introduce a text-imagination alignment loss to explicitly encourage visual grounding during training. We generate visual imaginations for key noun phrases in instructions that refer to visual landmarks using an off-the-shelf text-to-image generation model [32].

Applying our technique to two existing models, HAMT [8] and DUET [9], we find modest improvements – R2R validation unseen success increases by roughly 1.0 and 0.6 points respectively. This translates to an improvement of 2 points in success on the test set for our modified DUET model. On REVERIE validation unseen, we improve DUET performance by 1.3 points success and 0.82 points grounding success. Our results suggest that including explicit visual depictions of referred objects can improve performance on vision-and-language navigation tasks.

Contributions. We summarize our contributions as:

- We develop a pipeline for generating visual imaginations from navigation instructions and synthesize the R2R-Imagine dataset to enable studying the impact of text-to-image models on VLN agents.
- We propose an agent-agnostic method to incorporate visual imaginations into existing VLN agents and show improved performance for HAMT [8] and DUET [9] across R2R and REVERIE.
- We provide ablations of modeling decisions to characterize the design space for adding visual imaginations.

2. Related Work

Vision-and-Language Navigation (VLN). Navigating agents that can operate in novel environments using free-form natural language lends itself as personal assistants, field-robots and search-and-rescue agents. Development of VLN agents has been facilitated by photorealistic simulators such as Matterport3D [7] and paired language-action datasets such as Room-2-Room (R2R) [4] and REVERIE [33]. R2R provides fine-grained instructions conveying low-level directives such as "Go left, then right at the hall". REVERIE specifies coarse-grained instructions conveying broad high-level goals such as "Adjust the picture by the lamp in the hall". Progress in VLN has broadly ranged from architectural improvements and data scaling techniques. Architectural improvements include evolution from recurrent systems [4] to transformer architectures [8, 9, 17]. HAMT [8] uses a hierarchical transformer architecture to encode spatial and temporal information. DUET [9] maintains a topological map and performs coarse and fine scale encoding over it. Data scaling techniques include augmenting environments [14, 40] using an action-to-text speaker model [12] to generate synthetic instructions, and scaling environments [19, 41] with additional scenes such as HM3D [35] and Gibson [42] in Habitat [37]. Recently, zero-shot or learned integrations of pretrained large language models for VLN [45, 46] is showing promising performance. However, running large models in-the-loop incurs high computational costs.

Closely related to our work, ADAPT [27] augments VLN agents with visual observations related to an instruction. Language-observation pairs are stored from the training set and multiple pairs of matching object/location nouns from instructions are retrieved at test time using CLIP [34] scores. Fundamentally, the ADAPT image base is limited to training environments whereas our imaginations cover an open distribution. Therefore, our method can be more representative of the instruction and visualize novel objects/locations. For instance, a sub-instruction of "walk into the kitchen with blue walls" would retrieve images of kitchens from the training environment using ADAPT, but our imaginations would convey a kitchen with blue walls.

Another adjacent work, LAD [26] leverages layout dreamer and goal dreamer modules optimized around a topological graph to navigate via coarse-grained instructions. The capability of our method contrasts with LAD inthree primary ways: 1) method generality; Our method is performant across models and datasets, whereas LAD is a

full architecture constructed around a particular topological graph in coarse-grained settings, 2) sequential imagination: we imagine a sequence of landmarks along a path to inform sequential decision-making, whereas LAD imagines just the goal, and 3) we perform early fusion of imaginations and language whereas LAD employs late fusion.

Image generation in VLN. Several techniques have proposed generative models to predict future visual observations or entire environments in VLN [21, 24, 25]. Pathdreamer [21] developed a GAN-based model to predict panoramic observations after a navigation action in VLN, conditioned on prior visual observations. Agents utilizing these predicted observations in a fixed-horizon search with a learned trajectory-instruction alignment heuristic achieve non-trivial success compared to simple baselines, but fall short of planning on actual observations and standard VLN models. Unlike our approach, Pathdreamer does not condition on instruction text or explore training agents on generated imagery. Similarly, PanoGen [24] uses a diffusion model to also generate panorama observations, but conditions on image captions while doing so as a form of visual dataset augmentation – replacing Matterport3D observations with synthetic generations with matching semantics. In contrast to our approach, this data augmentation does not provide additional inputs to the VLN models, is not conditioned on instructions, and is not active at inference time.

Related to our work is VLN-Sig [25], which adapts VLN agents to predict quantized features of future visual observations given the instruction and observation history. This supports additional pretraining tasks and an auxiliary loss during downstream VLN finetuning for predicting next-step image semantics. Unlike our approach, VLN-Sig relies on the trained VLN agent to predict future image properties and focuses more on next-step prediction. In contrast, we leverage broad vision-and-language knowledge from text-to-image models and focus generating instruction-referred visual landmarks to improve agent performance.

Image generation in robot control. Taking a wider view, several methods for controlling embodied agents have leveraged text-conditioned generative techniques to provide goal specifications in the form of images or videos. For tabletop object rearrangement tasks, many works use text-to-image models to synthesize goals or sub-goals in response to natural language commands, see [6, 13, 20] for a non-exhaustive sample. Others produce denser conditioning, leveraging text-to-video models to generate entire potential robot trajectories [5, 11]. While similar in motivation to our approach, these techniques focus on tabletop manipulation settings which introduce difficult physical challenges but significantly limit the problem to constrained spaces that remain largely observed throughout operation. As such, sub-goals are often edits to existing observations rather than generations of yet-to-be-seen landmarks as in our approach.

3. Methodology

Our proposed methodology consists of two components – (1) a visual imagination generation pipeline and (2) a model-agnostic approach to integrating visual imaginations into existing Vision-and-Language Navigation (VLN) agents. Before defining these, we start with establishing relevant notation for the VLN task.

Problem definition. In the standard VLN task [4], an agent is prompted with a natural language navigation instruction and must make a sequence of actions to navigate along the described trajectory based on visual observations. We denote the L-word instruction as $W = (w_1, w_2, \cdots, w_L)$ and the visual observation at time step t as a panoramic observation O_t consisting of K single view images $O_t = (I_1^t, \cdots, I_K^t)$. Following prior work, we consider $K = 36$ corresponding to images from 12 headings and 3 pitches. Amongst the single view images, a subset $\mathcal{A}_t \subset O_t$ correspond to navigable directions. Given the instruction W and history of observations $O_{\leq t}$ until time t, the agent selects an action $a_t \in \mathcal{A}_t \cup \{\texttt{stop}\}$ and accordingly either moves or terminates the episode. Generally, the VLN agent is parameterized as a policy network $\pi_\theta(a_t \mid W, O_{\leq t})$ with parameters θ learned via general pretraining and a combination of imitation learning and reinforcement learning [8, 9].

3.1. Generating Visual Imaginations

To generate visual imaginations of potential landmarks, we consider the structure of navigation instructions. Typically, these consist of a sequence of sub-instructions describing intermediate navigation steps, which may or may not refer to visual landmarks. For instance, a sub-instruction "go past the couch" implies a visual landmark while "go straight then left" does not. We use FG-R2R [15] to segment instructions, resulting in an average of 3.66 sub-instructions per instruction for the R2R[4] training set. For a given instruction W decomposed into m segments, we denote these as sub-instructions $\mathcal{S} = (S_0, \cdots, S_m)$.

Sub-instruction filtering scheme. As sub-instructions may not always contain references to visual landmarks, we filter sub-instructions prior to image generation. First, we ignore any sub-instruction that lacks a noun phrase using Spacy [2] and then noun phrases are further filtered by a manual blacklist to avoid uninformative landmark generation. For example, we exclude detected noun phrases rooted on non-visual terms like counts ('one'), directions ('left'), or ambiguous object pronouns ("it") as these lack specificity to produce meaningful visual landmarks. After filtering, we denote the remaining sub-instructions $\mathcal{S}' \subset \mathcal{S}$.

Visual imagination generation. The filtered sub-instructions are passed through a text conditioned diffusion model as illustrated in Fig. 3. We use SDXL [32] as the diffusion model with positive and negative prompts cho-

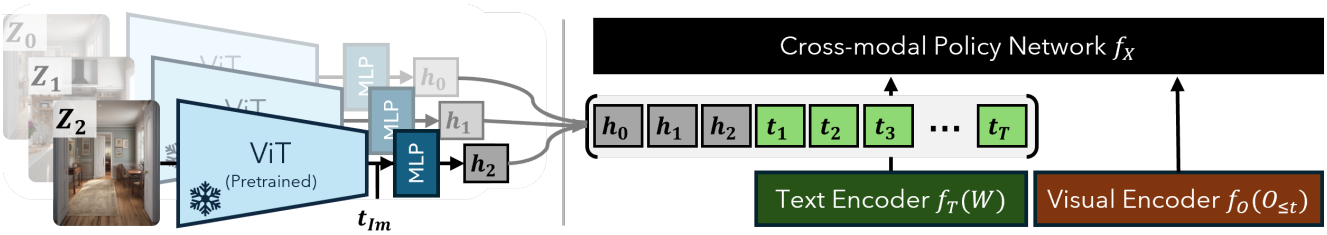


Figure 2. An overview of our approach. (Left) Imaginations generated using valid sub-instructions from an instruction as determined by our filtering scheme are first passed to a pre-trained ViT to obtain feature vectors. A type embedding t_{Im} for imagination modality is then added to the features which are encoded using a 3 layer MLP to obtain imagination embeddings h_i. (Right) To integrate imagination modality to a VLN agent, the imagination embeddings h_i are concatenated with instruction embeddings t_i that are encoded using a text encoder $f_T(W)$. The concatenated imagination-text embeddings are passed to the VLN agent's cross-modal encoder f_X along with visual embeddings to predict a distribution over the agent's action space.

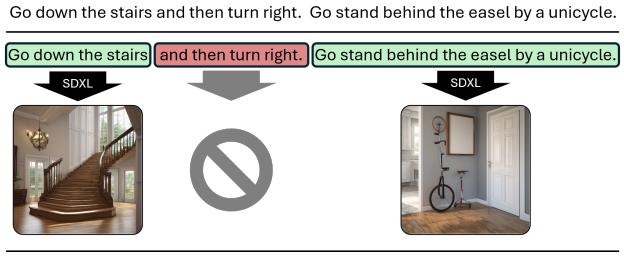


Figure 3. Example instruction segmentation, filtering, and image generation. Instructions are segmented to sub-instructions leveraging FG-R2R [15] and then filtered to remove phrases referring to uninformative nouns (e.g., "right"). We produce visual imaginations using SDXL [32] for remaining sub-instructions.

sen to steer generation towards indoor real-estate environments akin to typical environments in Matterport3D [7]. For example, providing "indoor" and "real estate" as positives prompts and "humans" and "collage" as negatives. See the supplementary for a full list. For a given instruction, we denote the generated images for valid sub-instructions as $\mathcal{Z} = \{ Z_i \mid S_i \in \mathcal{S}' \}$. We colloquially refer to these generated images as imaginations and generate them for the entire R2R dataset. The resulting R2R-Imagine dataset consists of over 41k synthesized imaginations. This amounts to an average of 2.96 imaginations per instruction.

3.2. Integration with Existing Models

We consider augmentation with visual imagination to be a model agnostic approach and integrate it in two prior models – HAMT [8] and DUET [9]. Our integration approach consists of an imagination encoder, an alignment-promoting auxiliary loss, and a finetuning regimen (Fig. 2). Without loss of generality, we consider a decomposition of any VLN agent into an instruction encoder $f_T(W)$, an observation/history encoder $f_O(O_{\leq t})$, and a cross-modal policy network $f_X(f_T(W), f_O(O_{\leq t}))$ that combines the encoder outputs to predict a distribution over actions. Note that these may be fairly complex mechanisms and include explicit history encoding modules separate from encoding the current observation; however, this abstraction is sufficient here.

Imagination encoder and integration. Given an instruction W with corresponding imaginations $\mathcal{Z}$, we independently encode each imagination Z_i using a pretrained vision transformer (ViT) [10] to produce a d-dim embedding vector. These are further transformed by a three-layer MLP such that the ith imagination embedding can be written as

$$h_i = \text{MLP}(\ \text{ViT}(Z_i) + t_{Im}\ ) \tag{1}$$

where t_{Im} is the type embedding for imagination modality. As we consider visual imaginations to be proxies for sub-instructions, the full set of imagination embeddings $\mathcal{H}$ are concatenated to the text encodings such that our general VLN agent structure now computes $f_X(\ [f_T(W), \mathcal{H}], f_O(O_{\leq t})\ )$ where $[\cdot,\cdot]$ denotes a concatenation. Taking the DUET model as an example, the imagination embeddings are concatenated to the instruction encoding prior to being passed to the coarse-scale and fine-scale cross-modal encoders.

Auxiliary alignment loss. We add an auxiliary loss to align the representations of imaginations and their corresponding sub-instruction's noun phrases. For a given sub-instruction S_i, we denote the mean embedding vector for noun phrase tokens in S_i from the text encoder f_T as $\bar{S}_i$. To align representations, we compute a cosine similarity loss between the corresponding imagination Z_i encoded as h_i and $\bar{S}_i$. This is averaged over all N_{Im} imagination-subinstruction pairs (Z_i, S_i) in a batch,

$$\mathcal{L}_{cos} = \frac{1}{N_{Im}} \sum_{(Z_i,S_i)\in B} \left(1 - \frac{h_i \cdot \bar{S}_i}{\|h_i\| \|\bar{S}_i\|}\right). \tag{2}$$

During finetuning, we combine this auxiliary loss $\mathcal{L}_{\text{cos}}$ with the losses from the base agent with a scaling factor λ such that the overall loss is $\mathcal{L}_{\text{base}} + \lambda \mathcal{L}_{\text{cos}}$.

Finetuning regimen. We start with fully-trained checkpoints of prior models as our initial weights. Depending

Table 1. Comparison of our approach with selected prior work on the R2R dataset. Methods that use additional visual data beyond MP3D [7] are annotated with † and are not directly comparable with other approaches. Note that BEVBert* additionally uses panoramic depth images. Adding our visual imagination approach to HAMT and DUET base models (gray rows) leads to improved success rate (SR) and success weighted by inverse path length (SPL) metrics on the val-unseen split. For DUET, this effect also persists on the test set. We highlight these improvements over base models using **bold** numbers.

Methods	Validation Seen				Validation Unseen				Test Unseen			
	TL	NE↓	SR↑	SPL↑	TL	NE↓	SR↑	SPL↑	TL	NE↓	SR↑	SPL↑
PREVALENT [14]	10.32	3.67	69	65	10.19	4.71	58	53	10.51	5.30	54	51
RecBERT [16]	11.13	2.90	72	68	12.01	3.93	63	57	12.35	4.09	63	57
ADAPT [27]	10.97	2.54	76	72	12.21	3.77	64	58	12.99	3.79	65	59
BEVBert* [3]	13.56	2.17	81	74	14.55	2.81	75	64	15.87	3.13	73	62
MARVAL [19]†	10.60	2.99	73	69	10.15	4.06	65	61	10.22	4.18	62	58
ScaleVLN [41]†	11.90	2.16	87	80	12.40	2.34	87	79	14.27	2.73	83	77
HAMT [8]	**11.15**	2.51	75.61	72.18	**11.46**	3.62	66.24	61.51	**12.27**	3.93	65	60
HAMT-Imagine (ours)	11.30	**2.36**	**77.16**	**73.72**	11.81	**3.58**	**67.26**	**62.02**	12.66	**3.89**	65	60
DUET [9]	**12.33**	2.28	78.84	72.88	**13.94**	3.31	71.52	60.41	**14.73**	3.65	69	59
DUET-Imagine (ours)	12.93	**2.19**	**79.9**	**73.75**	14.35	**3.19**	**72.12**	**60.48**	15.35	**3.52**	**71**	**60**

on the base method, these parameters may be the result of multi-stage pretraining and finetuning regimens – e.g., leveraging synthetic PREVALENT [14] data as prescribed in [8]. The imagination encoder is integrated to the base model and finetuned with the R2R-Imagine dataset. To mitigate catastrophic forgetting, we perform our training in three stages. First, we train only the newly introduced imagination encoder MLP along with type embeddings t^{Im} with the remaining parameters frozen. Then, all the modules are trained jointly with a decreased learning rate for the base model. Finally, all the parameters are trained at a common learning rate. See supplementary materials for details on learning rates and schedules.

4. Experiments

Using the R2R dataset, we evaluate the impact of visual imaginations on our agent's navigation ability, perform ablation studies to explore the design space, and illustrate qualitative impacts of visual imaginations. Using the REVERIE dataset, we study the generalizability of our technique to a fine-grained instruction setting.

4.1. Experimental setup

R2R dataset. R2R [4] is constructed using Matterport3D [7] simulator. It consists of 90 indoor environments with 10,567 panoramas (each panorama consists of 36 single-view images) which act as nodes in a navigation graph. There are 7,189 trajectories each with 3 instructions and a corresponding ground truth trajectory. The dataset is split into train (4675 trajectories), val-seen (340 trajectories), val-unseen (783 trajectories), and test (1391 trajectories). The train and val-seen split share environments while val-unseen and test have different environments.

R2R-Imagine dataset. The curation process of R2R-Imagine is prescribed in section 3.1. The number of imagination images are 41,558 for train, 3,055 for val-seen, 6,857 for val-unseen, and 12,412 for test. Our imaginations have a resolution of 1024x1024 and a single imagination generation on a single H100 GPU takes 3.2 seconds on average. We intend to release the dataset upon acceptance.

REVERIE dataset. REVERIE [33] is constructed using Matterport3D and consists of high-level goals instead of finer grained step-by-step instructions. These instructions generally convey the target object and its location. The agent has to rely more on exploration relative to R2R. To be successful, agents must reach and identify the referent object. To synthesize imaginations for REVERIE, we treat the entire instruction as a single sub-instruction and follow the same recipe used for R2R.

Evaluation metrics. We employ standard VLN metrics [18]. Success rate (SR) is the ratio of instruction-trajectories where the agent stops within a 3 meter radius from the goal and success normalized by inverse path length (SPL) helps distinguish successful episodes with shorter path lengths from longer path lengths. Navigation error (NE) is the distance between an agent's final position and the goal in meters. Trajectory length (TL) is the length of an agent's traversed path in meters. REVERIE [33] additionally defines grounding metrics Remote Grounding Success (RGS) and RGS penalized by path length (RGSPL).

Table 2. Performance comparison of our approach on REVERIE with baseline. We observe improvements of our method across all metrics on DUET when provided coarse-grained instructions.

	SR↑	SPL↑	RGS↑	RGSPL↑
DUET	46.98	33.73	32.15	23.03
DUET-Imagine (ours)	48.28	33.76	32.97	23.25

HAMT and DUET base models. We apply our approach to two baseline representative architectures – HAMT [8] and DUET [9]. HAMT is hierarchical and transformer-based; instructions, observations, and trajectory history are encoded separately using unimodal transformers and then combined in a single cross-modal transformer to predict actions. DUET incrementally forms a topographic map of the environment over time and uses dual fine- and coarse-grained cross-modal attention mechanisms to associate language embeddings with the local observation and global topographic features. The results are then fused to predict actions. For both base models, we inject visual imaginations via concatenation with the text embeddings (Sec. 3.2). We use provided checkpoints for pretrained models.

Implementation details. We follow base agent [8, 9] settings unless specified otherwise. We use a standard off-the-shelf pretrained ViT-B/16 [10] to encode imaginations for both HAMT and DUET. Our VLN-Imagine agents are finetuned for 100k iterations in three stages as noted in section 3.2. The finetuning process is carried out using a Tesla V100 GPU with a batch size of 8 and takes $\sim$ 1.5 days for each agent. For combining the proposed auxiliary loss function with the base model losses, we empirically set λ=0.5.

4.2. Results

Our experiments demonstrate that providing generated images of instruction landmarks (*imaginations*) improves the performance of vision-and-language navigation models (Tab. 1). We highlight our key findings below.

VLN agents perform better with imaginations. VLN agents equipped with imaginations perform better than VLN agents without by 0.6-1.0 SR and up to 0.5 SPL on R2R val-unseen (Tab. 1). This improvement coincides with lower navigation error (NE) across both models and all dataset splits. We contextualize these results relative to agent architectures that are representative of the VLN literature. We additionally report a distinct family of techniques that scale training environments to boost performance by reducing overfitting – MARVAL [19] and ScaleVLN [41]. We consider approaches of this class orthogonal to our study and instead use visual data only from Matterport3D [7].

Imaginations are useful in coarse-grained instruction navigation. Our method is performant when natural language instructions convey high-level objectives (REVERIE, Tab. 2); navigation performance improves by 1.3 SR and grounding performance improves by 0.82 RGS. We hypothesize that imaginations not only aid in reaching the goal but also in identifying and grounding the target concept amongst others objects in the scene.

Table 3. Role of imaginations; HAMT-Imagine. Null imaginations refers to testing with no imaginations, wrong imaginations refers to testing with imaginations sampled from different instructions. We observe aligned imaginations are important to navigation performance. In addition, results with no imaginations are better than baseline implying a regularization effect during training.

	SR↑	SPL↑
Baseline (HAMT)	66.24	61.51
Null imaginations	66.92	61.89
Wrong imaginations	66.24	61.02
Correct imaginations (ours)	67.26	62.02

Table 4. Sequential *vs*. goal imagination; HAMT-Imagine. We study the effect of providing just the imagination belonging to the final valid sub-instruction relative to full valid imaginations. We observe providing just the goal imagination improves performance significantly compared to baseline but shows lower performance relative to providing full imaginations.

	SR↑	SPL↑
Baseline (HAMT)	66.24	61.51
Goal imagination	66.79	61.58
Full imaginations (ours)	67.26	62.02

VLN agents benefit from both imagination *training* and *inference*. When nullifying imaginations at test-time, VLN-Imagine agents still surpass baseline performance, but to a lesser extent (Tab. 3). We implement this nullification by setting the imagination attention masks to zero in the cross-modal encoder. We hypothesize that imagination-based training provides a regularizing effect on the overall VLN agent during fine-tuning, revealing a research avenue for strong and performant test-time model inference. We further find that imaginations must be aligned with the instruction to provide maximal benefit; when fed with imaginations randomly-sampled from the R2R-Imagine dataset, VLN-Imagine agents perform worse than baseline (Tab. 3).

Sequential imagination outperforms goal imagination. Sequential refers to imaginations generated over multiple sub-instructions (our method). Goal imagination refers to imagining just the final non-filtered sub-instruction. We show these results in Tab. 4 – while goal imagination outperforms the baseline by 0.5 SR, sequential imagination outperforms the baseline by 1.0 SR. Goal imagination may be particularly useful in guiding the agent to stop at the destination. While distinct from our setting, such goal imagi-

Table 5. Comparison between ViT models. We train both HAMT-Imagine and DUET-Imagine agents with a ViT finetuned in-the-loop with HAMT and an off-the-shelf pretrained ViT. We observe no consistent effects across the both agents.

	ViT fine-tuned		ViT off-the-shelf	
	SR↑	SPL↑	SR↑	SPL↑
HAMT-Imagine	67.31	61.5	67.26	62.02
DUET-Imagine	72.58	60.29	72.12	60.48

Table 6. Loss ablation. We study the effect of negative samples in our alignment loss by comparing our approach of using cosine loss with InfoNCE [28] loss. In addition, we train the agent with no auxiliary loss represented by "No loss". We do not observe a significant difference between contrastive and cosine losses.

	SR↑	SPL↑
No loss	66.75	61.60
$\mathcal{L}_{\text{InfoNCE}}$	67.18	61.84
$\mathcal{L}_{\text{cos}}$ (ours)	67.26	62.02

nation has parallels to ImageNav [22], a visual navigation task that conveys goals solely using exemplar images.

General-purpose vision encoders are sufficient imagination encoders. We compare encoding imaginations with off-the-shelf ViT [10] *vs*. ViT fine-tuned on R2R navigation episodes [8]. We find comparable performance (Tab. 5) and choose to employ off-the-shelf ViT in our final model. This is motivated to maximize generality and to avoid potential issues relating to catastrophic forgetting. For example, generalization of frozen ViT is demonstrated in Zeng et al. [44].

Aligning imaginations to instructions benefits VLN-imagine agents. Aligning the imagination embedding to the instruction embedding as an auxiliary loss leads to a 0.5 SR and a 0.4 SPL increase over no auxiliary loss (Tab. 6). We also consider formulating this loss as contrastive – do negative samples enable a stronger alignment and thus better downstream performance? We find that an InfoNCE loss [28] yields no significant difference relative to our cosine similarity loss (Eq. 2). We construct negative samples using mean noun phrase embeddings from imaginations of other instructions in the batch. The overall loss is $\mathcal{L}_{\text{base}} + \lambda\mathcal{L}_{\text{InfoNCE}}$ where we set λ to 0.2 empirically.

Imaginations correctly represent noun phrases in a sub-instruction. To evaluate fidelity of our imaginations, we run an open-vocabulary object detector (LangSAM [1]) on our imaginations to test if the noun phrases contained in the corresponding sub-instruction from R2R are detected successfully. We see in Tab. 7 that at least one noun phrase is detected in our imaginations for 98.78% of sub-instructions and all noun phrases are detected for 94.99% of sub-instructions. We conclude our imaginations provide a high coverage of relevant concepts across splits.

Table 7. Open vocabulary object detection accuracy (%) of imaginations to noun phrases from sub-instructions. Our imaginations represent the noun phrases with a high accuracy. In val-seen, in 98.43% of sub-instructions, at least 1 noun phrase is detected and in 94.51% of sub-instructions, all noun phrases are detected.

	Sub-instructions	Noun Phrase Detection (%)	
		≥1 detected	All detected
Val-seen	2985	98.43	94.51
Val-unseen	6700	98.99	95.34
Test	12075	98.93	95.12

4.3. Qualitative visualizations

In this section, we provide a qualitative investigation of how imaginations may be used as a pivot between language tokens and visual observations from the environment by examining the attention patterns in HAMT-Imagine.

For a given instruction, we randomly select an imagination / sub-instruction pair and examine the agent's trajectory to identify the first occurrence of the reference object. We then examine the attention distributions between language tokens, the visual imagination, and the panoramic observation. As there are many attention heads across multiple layers in the cross-modal encoder, we structure our study by focusing on those attention heads where sub-instruction tokens are attended to most strongly by their corresponding visual imagination and then examine how the visual imagination attends to observations.

We then visualize sub-instruction and imagination pairs along with their top attended language tokens and visual observations in Fig. 4. By construction, these are time steps where the referent is visible and the visual imagination is attending to the appropriate noun phrase. As such, our examples primarily demonstrate how the imagination learns to associate with the visual observations. We focus on this aspect as our auxiliary loss already explicitly encourages alignment between visual imaginations and sub-instruction.

From the examples in Fig. 4, we make two remarks. *Imaginations attend most strongly to observations of matching semantic concepts*. In row 1, the imagination and the retrieved observations all depict a cheetah (either a painting or floor mat). *Imagination attention scores mirror the proper disambiguation of concepts*. In row 2, the sub-instruction (and thus the imagination) refer to a unicycle. While both a bicycle and unicycle are observed during execution, attention scores are maximized by the observations containing the unicycle. These patterns do not directly imply a causal relation between the attention scores and eventual decision making of the VLN agent. Rather, they illustrate how imag-

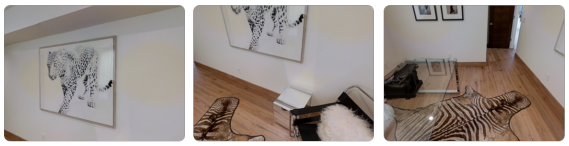
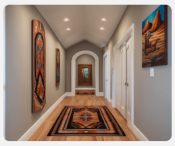
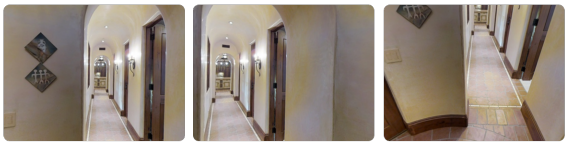
Sub-instruction	Imagination	Top Attended Tokens	Top Attended Observations
And walk towards the door pass by the cheetah painting		che, ##eta, ##h, painting	
Go stand behind the easel by a unicycle next to a door		un, ##ic. ##y, ##cle, to	
Once in the room, walk to the left of the pool table		pool, door, room	
Walk into the arch hallway near the Native American painting on the wall		American, Native, candles	

Figure 4. Qualitative examples showing imaginations as pivots between language and observation images. The first column contains sub-instruction from a random instruction from R2R, the second column contains the imagination generated using the sub-instruction. The third and fourth columns show highest attended language tokens and observation images from an attention head in HAMT's cross-modal transformer at a time step the associated observations are first visible. In the second example (row 2), the sub-instruction references "unicycle" which is captured in the imagination along with neighboring nouns "easel" and "door". We observe that in a head where top attending language tokens to the imagination query are references to nouns associated with the sub-instruction, its top attended observations to the imagination query are images of the same concept ("unicycle"). In this example, the imagination of a unicycle is being used to associate language tokens belonging to "unicycle" to observations of unicycle hinting at the utility of imaginations in navigation.

inations can serve as a pivot between language and observations and afford disambiguation of related concepts.

5. Conclusion

In this work, we show that visual representations of instructions, or visual imaginations, play an additive role in training superior navigating agents for vision-and-language navigation (VLN) tasks. We generate a sequence of visual imaginations depicting sub-instruction landmarks using a text-to-image diffusion model. Our approach is model-agnostic and we integrate these imaginations into VLN agents using an imagination encoder, an auxiliary loss, and a fine-tuning training step. We apply this approach to two representative VLN agents across two datasets and observe performance improvements of around 1 SR and 0.5 SPL.

Limitations. Generating and encoding imaginations increases the computational cost of running VLN agents. This is particularly relevant when running agents on-device for real-world robotic deployments. However, we note that this process is performed upfront, unlike techniques that depend on step-wise observations like novel view synthesis (NVS). Stepping more broadly, by design, our imaginations are not grounded in the environment. Personalized reasoning, such as the unique naming of objects and locations, cannot be directly imagined in our framework. Life-long learning of persistent groundings is an open question not just in generating imaginations, but in the VLN space at large.

Future directions. A natural question that arises from our research is the exact complementary nature of the roles played by visual imaginations and language instructions. This is a rich area for future work. Similar studies relating language and visual observations have been conducted in [48] and [47]. Our paper demonstrated the fundamental utility of imagination in VLN agents. Beyond this lies interesting extensions, such as how imagination can help bridge the simulation-to-reality (Sim2Real) gap for VLN agents and whether imaginations can unlock the performance of VLN world models through image-image reasoning.

Acknowledgments. We thank Zijiao Yang and Robert Yelle for their assistance with simulator issues and technical support for high performance computing infrastructure.

References

[1] Language segment-anything. https://github.com/luca-medeiros/lang-segment-anything/. 7

[2] spacy: Industrial-strength natural language processing in python. https://spacy.io/. 3

[3] Dong An, Yuankai Qi, Yangguang Li, Yan Huang, Liang Wang, Tieniu Tan, and Jing Shao. Bevbert: Multimodal map pre-training for language-guided navigation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 5

[4] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2017. 2, 3, 5

[5] Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024. 3

[6] Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Rich Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pre-trained image-editing diffusion models. In *The Twelfth International Conference on Learning Representations*. 3

[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *2017 International Conference on 3D Vision (3DV)*, pages 667–676, 2017. 2, 4, 5, 6

[8] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. In *NeurIPS*, 2021. 2, 3, 4, 5, 6, 7

[9] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *CVPR*, 2022. 2, 3, 4, 5, 6

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4, 6, 7

[11] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[12] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. In *Neural Information Processing Systems (NeurIPS)*, 2018. 2

[13] Jialu Gao, Kaizhe Hu, Guowei Xu, and Huazhe Xu. Can pre-trained text-to-image models generate visual goals for reinforcement learning? *Advances in Neural Information Processing Systems*, 36:38297–38310, 2023. 3

[14] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training, 2020. 2, 5

[15] Yicong Hong, Cristian Rodriguez, Qi Wu, and Stephen Gould. Sub-instruction aware vision-and-language navigation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3360–3376, 2020. 3, 4

[16] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1643–1653, 2021. 5

[17] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1643–1653, 2021. 2

[18] Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. General evaluation for instruction conditioned navigation using dynamic time warping. *ArXiv*, abs/1907.05446, 2019. 5

[19] Aishwarya Kamath, Peter Anderson, Su Wang, Jing Yu Koh, Alexander Ku, Austin Waters, Yinfei Yang, Jason Baldridge, and Zarana Parekh. A new path: Scaling vision-and-language navigation with synthetic instructions and imitation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10813–10823, 2023. 2, 5, 6

[20] Ivan Kapelyukh, Vitalis Vosylius, and Edward Johns. Dall-e-bot: Introducing web-scale diffusion models to robotics. 2023. 3

[21] Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Pathdreamer: A world model for indoor navigation, 2021. 3

[22] Jacob Krantz, Stefan Lee, Jitendra Malik, Dhruv Batra, and Devendra Singh Chaplot. Instance-specific image goal navigation: Training embodied agents to find object instances, 2022. 7

[23] Chongxi Lai, Shinsuke Tanaka, Timothy D. Harris, and Albert K. Lee. Volitional activation of remote place representations with a hippocampal brain–machine interface. *Science*, 382(6670):566–573, 2023. 2

[24] Jialu Li and Mohit Bansal. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. *arxiv*, 2023. 3

[25] Jialu Li and Mohit Bansal. Improving vision-and-language navigation by generating future-view image semantics. *CVPR*, 2023. 3

[26] Mingxiao Li, Zehao Wang, Tinne Tuytelaars, and Marie-Francine Moens. Layout-aware dreamer for embodied visual referring expression grounding. In *Proceedings of the*

AAAI Conference on Artificial Intelligence, pages 1386–1395, 2023. 2

[27] Bingqian Lin, Yi Zhu, Zicong Chen, Xiwen Liang, Jianzhuang Liu, and Xiaodan Liang. Adapt: Vision-language navigation with modality-aligned action prompts. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15375–15385, 2022. 2, 5

[28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 7

[29] Allan Paivio. Abstractness, imagery, and meaningfulness in paired-associate learning. *Journal of verbal learning and verbal behavior.*, 4(1), 1965-2. 2

[30] Allan Paivio. A factor-analytic study of word attributes and verbal learning. *Journal of verbal learning and verbal behavior.*, 7(1), 1968-2.

[31] A. Paivio, P. C. Smythe, and J. C. Yuille. Imagery versus meaningfulness of nouns in paired-associate learning. *Canadian Journal of Psychology / Revue Canadienne De Psychologie*, 22:427–441, 1968. 2

[32] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 2, 3, 4

[33] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020. 2, 5

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[35] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021. 2

[36] Marianne Cumella Reddan, Tor Dessart Wager, and Daniela Schiller. Attenuating neural threat expression with imagination. *Neuron*, 100(4):994–1005.e4, 2018. 2

[37] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019. 2

[38] Daniel L. Schacter, Donna Rose Addis, and Randy L. Buckner. Remembering the past to imagine the future: the prospective brain. *Nature Reviews Neuroscience*, 8(9):657–661, 2007. 2

[39] Roger N. Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971. 2

[40] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2610–2621, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 2

[41] Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Hao Tan, and Yu Qiao. Scaling data generation in vision-and-language navigation. In *ICCV 2023*, 2023. 2, 5, 6

[42] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018. 2

[43] Zijiao Yang, Arjun Majumdar, and Stefan Lee. Behavioral analysis of vision-and-language navigation agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2023. 1

[44] Kuo-Hao Zeng, Zichen Zhang, Kiana Ehsani, Rose Hendrix, Jordi Salvador, Alvaro Herrasti, Ross Girshick, Aniruddha Kembhavi, and Luca Weihs. Poliformer: Scaling on-policy rl with transformers results in masterful navigators. In *8th Annual Conference on Robot Learning*, 2024. 7

[45] Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7):7641–7649, 2024. 2

[46] Gengze Zhou, Yicong Hong, Zun Wang, Xin Eric Wang, and Qi Wu. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In *European Conference on Computer Vision*, pages 260–278. Springer, 2025. 2

[47] Wanrong Zhu, Yuankai Qi, Pradyumna Narayana, Kazoo Sone, Sugato Basu, Xin Wang, Qi Wu, Miguel Eckstein, and William Yang Wang. Diagnosing vision-and-language navigation: What really matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5981–5993, 2022. 1, 8

[48] Wang Zhu, Ishika Singh, Yuan Huang, Robin Jia, and Jesse Thomason. Does vln pretraining work with nonsensical or irrelevant instructions? *ArXiv*, abs/2311.17280, 2023. 8