# Supplementary Material of Hijacking Vision-and-Language Navigation Agents with Adversarial Environmental Attacks

## 1. Data Description

### 1.1. R2R Data

As described in Sec 3.2, we select attack instances from `R2R-val-seen` that have sufficient support from `R2R-train`. In total, we generated 1735 episodes as `Train`, 577 episodes as `Validation` that corresponds to 273 attack instances as `Test`. The attack instances cover 68 unique objects spanning 39 environments. Notice the attack for each attack instance trains independently, so each attack instance has their corresponding train, validation set. Here we provide the aggregated number.

### 1.2. RxR Data

As attacks on RxR data taking substantially more compute and time compared to R2R, we random sample a subset based on the number of unique attack objects involved, which results in a subset covering 20 unique objects residing in 9 environments. As a result, we have 1659 episodes for `Train`, 345 for `Validation` that corresponds to 254 attack instances in `Test`. This is comparable to the number of attack instances of R2R.

### 1.3. Ablation Data

Similarly, to accommodate for time and compute constraints, we random sample a subset from R2R, that covers 34 unique objects out of 68 in total, which spans across 27 environments and result in 955 `Train`, 306 `Validation` and 147 `Test` attack instances, that is roughly half of the total dataset on which we reported main result.

## 2. lmer Construction for Factor Analysis

We investigate the effects from different factors on trajectory-level attack effectiveness on R2R `Test`. To facilitate our analysis, we frame experiments as paired-measurements on individual attack instances with nDTW measured pre- and post-attack. Let $Y_{hijk}$ be the response variable nDTW, where $j, k$ respectively index random effect grouping factors for individual objects $object \sim N(0, \sigma_o^2)$ and attack instances $instance \sim N(0, \sigma_s^2)$. We assess the statistical significance of some $predictor$ (e.g., object size) indexed by $i$ with $n$ levels affecting nDTW by fitting linear mixed effect regression (`lmer`) models of the form:

$$Y_{hijk} = \beta_0 + \beta_1 attack_h + \sum_{m=2}^{n} \beta_m (attack_h \times predictor_i)_m + object_j + instance_k + \epsilon_{hijk}, \quad (1)$$

where $\epsilon_{hijk} \sim N(0,1)$ is a random error term, $\beta_0$ the model intercept, $\beta_1$ the fixed effect of $attack$, and $\beta_{2:n}$ the fixed effect of the interactions between $attack$ and $predictor$. Note that there are exactly two samples for each $instance_k$, one for pre-attack ($attack_{h=0}$) and one for post-attack ($attack_{h=1}$). Informally, we model paired nDTW measurements as a main effect from applying the adversarial attack, the interaction effect between the attack and some predictor, and random intercept effects from attack instances and objects. We use an ANOVA to determine the overall significance of factors and examine graphical model residual diagnostics to validate its modeling assumptions. For factors with significant effects, we use a post-hoc t-test to determine if the coefficients relating to post-attack $predictor$ interactions are significantly different from zero. That is, we verify that the difference in effect from the $predictor$ in pre/post-attack measurements is significant, and that the strength of that effect in the post-attack setting is significantly different from zero.