REASEACH IDEA

**By Xingxin Yang**

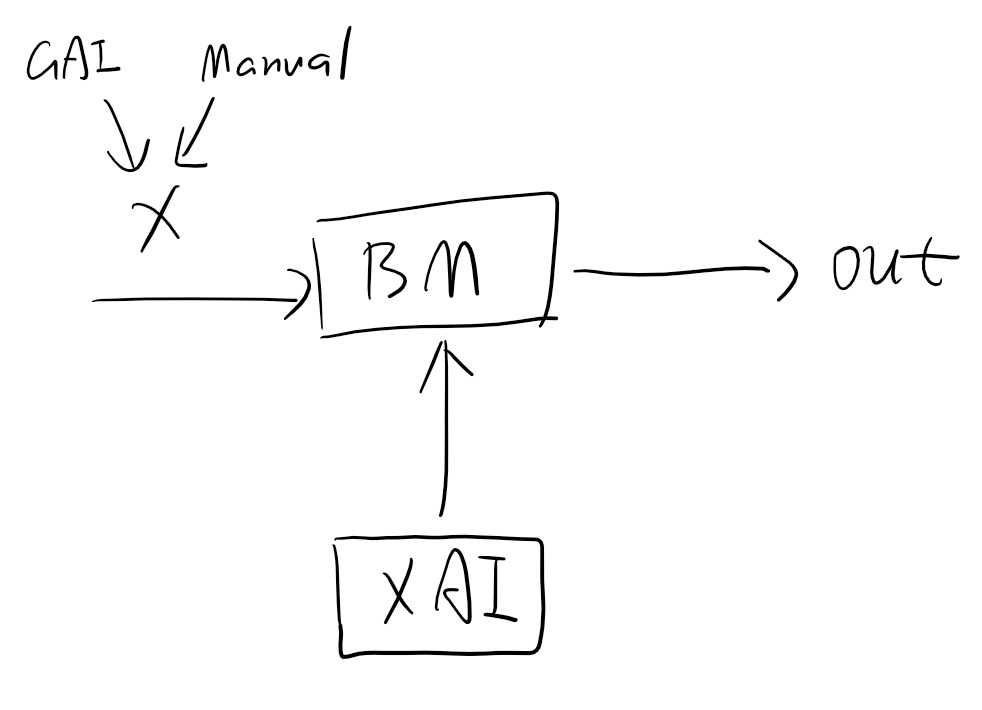
**March 28, 2024**

**Highlights:**

1. **Unify the evaluations of XAI by the perspective of destitution distance;**
2. **Using stable diffusion models to generate test samples with elaborate purpose;**
3. **Evaluating the fairness of XAI from different skins, genders, disability, ages;**
4. **Search-based methods, evaluating the XAI from multiple dimensions.**

**Details:**

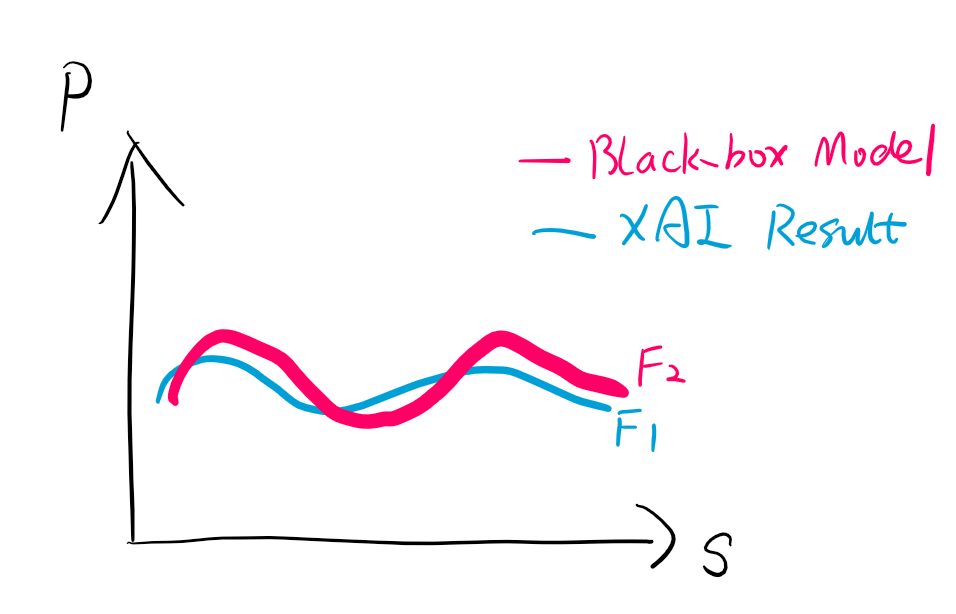
**An XAI problem can be illustrated as follow:**



**Where, X is the input data which can be generated by generative AI models such as diffusion network, or by hand. While BM is the short of Black-box Model. In this paper, we consider the segment anything model (SAM), SSD and Yolo models as BM for the reason that they are heavily related to the essential realms such as auto driving and medicine. (Although SAM, SSD and Yolo is opensource, we still regard them as BMs)**

**Meanwhile, serval well-known XAI methods such as LIME, Grad-CAM and SHAP are utilized to test the interpretability of BMs.**

**In order to test the performance of popular XAI methods, we redefine the problem as the distance of two distributions. Here:**



**The objective of testing problem can be described as optimization probelm:**

**Minimize: Distance(F1, F2)**

**It’s hard to formula the F2 mathematically in spite of the explainable F1. So this paper conduct the optimization process by trails. According to the law of large numbers, F1 and F2 can be calculated by a great number of trails. For a trail, the data *x* input to the *BM*, and output a result *out*. Then we run the XAI method to explain the relationship between the output and the feature maps inside the model, marked as *e.* With different trails, a set would be acquisited as follows:**

**{*xi*, *outi*, *ei*}, *i belongs to N, the number of trails.***

**Experiments:**

1. **Locally modify**

**E1) Prompt-based,**

**反证法**

**making diverse prompts to generate the suitable x and changes.**

**Trail types:**

1. **Keep x constant, change BMs’ layers, then make comparisons between e and out.**
2. **Keep out constant, change x, then make comparisons between *ei*-1 and *ei*.**
3. **Keep x and out constant, change BMs’ layers, then make comparisons between *ei*-1 and *ei*.**

**归纳法**

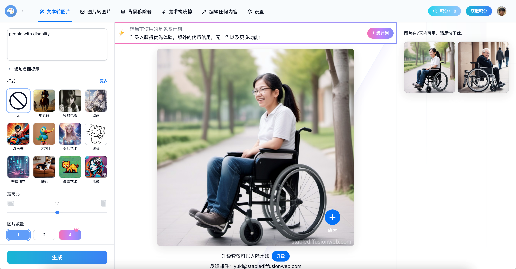
**E2) Search-based, generating various inputs, test the BMs and evaluate XAI methods, then summarizing the regular (i.e. The distributions of BMs and XAI), finally calculating the distance of F1 and F2.**

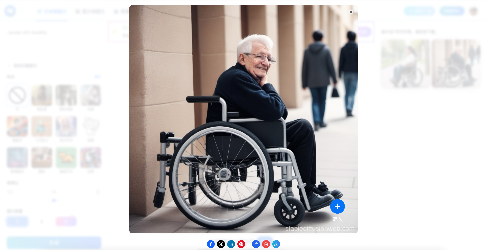
**The performance or the metrics of explainability *e* can be divided into some important aspects, such as fairness, failfulness, robustness. Here, we mainly take fairness into account.**

**For E1) Prompt-based method, we generated the inputs by stable diffusion models. SORA**

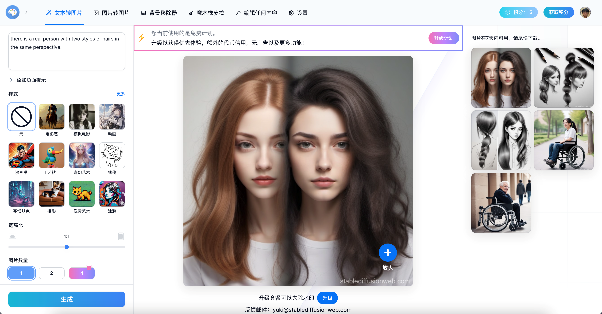
**Examples:**

**Prompt:** **people with disability**

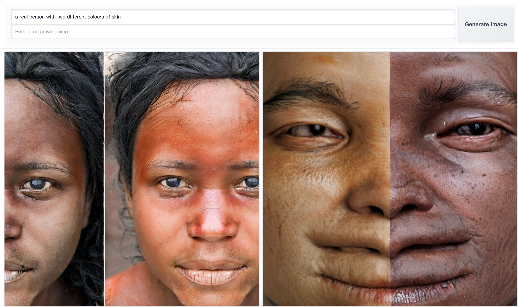
****

****

**Prompt:** **there is a real person with two styles of hairs in the same perspective.**

****

**Prompt: a real person with two different colors of skin**

****

****

****

**Appendix:**

**Test GAI**

**Grad CAM:**

