# Review 135: [Short] ORES: Open-vocabulary Responsible Visual Synthesis, 01.09.23 https://arxiv.org/abs/2308.13785.pdf

**Paper: https://arxiv.org/abs/2308.13785v1**

כיום מודלים לגנרוט תמונות לפי תיאור טקסטואלי מסוגלים ליצור תמונות באיכות מדהימה שממש תואמות את התיאור. בנוסף נהיה קשה מאוד עד בלתי אפשרי להבחין אם תמונה הינה טבעית״ נוצרה על ידי מודל גנרטיבי שפותח פתח ליצירת פייקים באיכות גבוהה מאוד שעלולים לגרום לנזק רב.  
  
היום ב-#shorthebrewpapereviews סוקרים מאמר שפיתח שיטה פשוטה למניעת יצירת פייקים מסוכנים. השיטה המוצעת מאוד פשוטה אינטואיטיבית. נניח שיש לנו פרומפט שבאמצעותו המשתמש רוצה ליצור תמונה. בנוסף יש לנו גם יש לנו גם סט של קונספטים (מושגים) אסורים (כמו ערום, דם וכאלו). בשלב הראשון הופכים את הפרופמט לבטוח (בהתאם לקונספטים האסורים) עם מודל שפה מאומן.   
  
המחברים בחרו דאטהסט קטן המכיל שלישיות של (פרומפט, מושג אסור, פרומפט בטוח) ובהתבסס עליו מצאו את הפרומפט (אחד) למודל שפה שבאמצעותו ניתן להפוך הנחיה נתונה יחד עם הקונספט האסור להנחיה בטוחה על ידי הרצה של הדאטהסט הזה על מודל שפה מספר אפוקים. לאחר מכן מכניסים את הפרומפט בטוח למודל דיפוזיה מאומן. כדי לגנרט תמונה דומה לפרומפט המקורי ב- S הצעדים ההתחלתיים של מודל הדיפוזיה (מתחילים מרעש טהור) וב- T – S הצעדים האחרונים מבצעים עם הפרומפט הבטוח.