# Review 158: [Short] Linguistic Binding in Diffusion Models: Enhancing Attribute Correspondence through Attention Map Alignment

**Paper: https://arxiv.org/abs/2306.08877v3**

https://arxiv.org/abs/2306.08877

מודלי דיפוזיה מודרניים מצטיינים ביצירת תמונות באיכות מרהיבה מתיאור טקסטואלי (ובטח DALLe3) וברוב המקרים התמונה ממש מתאימה לתיאור. אולם עדיין יש מקרים שמודל מתבלבל למשל בין הצבעים של האובייקטים המופיעים בתיאור. היום ב-#shorthebrewpapereviews סוקרים מאמר 🇮🇱 המציע שיטה למניעת בלבול סמנטי בין תכונות האובייקטים בתמונה.

הגישה המוצעת הינה פשוטה ואלגנטית. בשלב הראשון המחברים בונים את גרף התלויות הסינטקטית של הפרופמט כלומר מפיקים את כל קבוצות המילים (נגיד שם עצם ושם תואר) המתאימים אחד לשני (כמו (ארנב, צהוב) או (כורסא, בסגנון, מלון). לאחר מכן המחברים מכיילים מודל שפה עם פונקציה לוס ש״מפקחת״ על הדיוק הסמנטי של האובייקטים בתמונה.

איך זה נעשה? אתם בטח יודעים שייצוג של כל מילה בפרומפט (מופק באמצעות מודל שפה) מוזן למודל דיפוזיה. אז פונקציית לוס זו מנסה לאכוף דמיון בין מפות ה-attention בין מילים מאותה קבוצת שייכות סמנטית (שנבנה בשלב הקודם). מפות ה-attention אלו הם למעשה ציון cross-attention בין ייצוגי המילים (טוקנים) לבין פאטצ'ים בתמונה.

איבר נוסף בפונקצית הלוס מכיל איבר הממקסם מרחק בין מפות ה-attention בין המילים שלא שייכים לאותה קבוצה. הדמיון בין מפות ה-attention מחושב באמצעות מרחק KL סימטרי (נקרא גם JSD). המחברים טוענים כי הלוס זה מופעל בחלק מהאיטרציות של מודל דיפוזיה - על החצי הראשון של האיטרציות.