# Review 123: [Short] TextDiffuser: Diffusion Models as Text Painters, 13.08.23 https://huggingface.co/papers/2305.10855

**Paper: https://arxiv.org/abs/2305.10855v5**

מודלי דיפוזיה מצטיינים ביצירת תמונות מרשימות מתיאור טקסטואלי אך עדיין מתקשות ביצירה תמונות המכילות טקסט כחלק מהתמונה. למשל יצירת תמונה של כלב המחזיק שלט שכתוב עליו ״ברוך הבא הביתה״ עלולה ליצור תמונה עם טקסט שונה ולא ברור על השלט.

היום ב #shorthebrewpapereviews סוקרים מאמר שמנסה לתת מענה לסוגיה זו. המאמר מציע גישה דו-שלבית שבשלב הראשון נוצרת תמונה שבה נוצר את החלק בתמונה המכיל טקסט ובשלב השני מלבישים על התמונה זו את האובייקטים שיש בתמונה ושאר הפרטים (כגון טקסטורה ורקע). בשלב הראשון קודם כל בונים את שיכון (embedding) של הטקסט עם CLIP מאומן.   
  
אך להבדיל ממודלי דיפוזיה גנרטיביים אחרים מוסיפים לוקטור השיכון מוסיפים לכך שיכון נלמד של מילות המפתח (מחלקים את התיאור למילים שצריכות להופיע בתמונה ואלה שלא ובונים וקטורי שיכון שלהם). בנוסף מוסיפים לוקטור השיכון קידוד נלמד של רוחב של כל אות בתמונה ובנוסף מוסיפים לכך קידוד תלויה מיקום (positional encoding) נלמד. כל השיכונים הללו מחושבים באמצעות שני טרנספורמרים: אנקודר ודקודר.  
  
הראשון מאומן לקודד את הדאטה, השני מחשב) (Bounding Boxes (BB עבור האותיות בתמונה ובשלב האחרון מרנדרים את התמונה לפי ה-BB שחישבנו והאותיות (המקודדות) ומגנרטים מסכות לאותיות. בשלב השני קודם כל מגנרטים תמונה כאשר הקלט הוא מסכות הנוצרות בשלב הראשון (בכמה צורות), השיכון של הטקסט והתמונה המורעשת (הרי זה מודל דיפוזיה). בנוסף ללוס הרגיל של מודל הדיפוזיה המודל נקנס על אי התאמה של מיקום האותיות בתמונה (הם מאמנים רשת לזיהוי מיקומים אלה).