# Review 122: [Short] DIVIDE & BIND YOUR ATTENTION FOR IMPROVED GENERATIVE SEMANTIC NURSING, 12.08.23 https://huggingface.co/papers/2307.10864

**Paper: https://arxiv.org/abs/2307.10864v3**

בטח שמתם לב שלפעמים שאתם מבקשים ממודל גנרטיבי (נגיד MidJourney) ליצור תמונה עם כמה אובייקטים (נגיד חתול, כלב, שולחן ומקרר) התמונה לא תמיד מכילה את כל האובייקטים במיוחד כאשר התיאור הוא די ארוך ומכיל מספר רב של אובייקטים.   
  
היום ב-#shorthebrewpapereviews סוקרים מאמר המציע מענה לסוגיה הזו. קודם כל נבין למה לא תמיד אנו מצליחים להעביר למודל גנרטיבי (מודל דיפוזיה) את כל המידע. כדי להבין זאת נציין שהמידע (ייצוג) של הטוקנים מוזן למנגנון cross-attention לתוך השכבות הפנימיות של רשת UNet שהיא הלב של מודל הדיפוזיה (משערכת את הרעש בכל איטרציה).   
  
אז לפעמים טוקן טקסטואלי אחד ״גונב״ את כל ״תשומת הלב״ ואז הטוקנים האחרים פשוט לא באים לידי ביטוי ונעלמים מהתמונה. סוגיה נוספת שעלולה לקרות כאן היא attribute binding שבו פיצ'רים של אובייקטים מסוימים (כגון צבע או טקסטורה) משויכים לאובייקטים אחרים. המחברים מציעים להתמודד עם סוגיות אלו עם שתי גישות חדשות ל״הזזה״ של ייצוג הדאטה בכל איטרציה (semantic guiding) לכיוון של גרדיאנט הפונקציה המנסה לאכוף תכונות רצויות של מפות ה-attention. קודם כל הוא מנסה לכפות שונות מקסימלית ביו מפות ה-attention של כל אובייקט בין טוקנים ויזואליים סמוכים (פאצ'ים של תמונה). לטענת המחברים ככה מונעים מאובייקט אחד להשתלט לנו על כל הטוקנים הויזואליים.   
  
התכונה השניה שמנסים לאכוף היא שוני בין מפות attention של אובייקטים שונים. את זה הם משיגים עם מקסום של מרחק (Jensen-Shannon (JSD בין מפות ה-attention המנורמלות בין כל שני האובייקטים. כאמור מזיזים את שערוך הייצוג של כל איטרציה של מודל דיפוזיה בכיוון שסכום מנורמל את שני ה״יעדים״ האלו.