# Review 86: [Short] BuboGPT: Enabling Visual Grounding in Multi-Modal LLMs

**Paper: https://arxiv.org/abs/2307.08581v1**

רוצים מודל העונה על השאלות לגבי האובייקטים בתמונה וגם מראה לכם איפה האובייקט נמצא בתמונה ולתת הסבר לגביו? בנוסף לאודיו נתון אתו המודל יותר לסמן אותכם מי בתמונה שלכם הוא מקור האודיו. וכל זה מלווה בהסברים

היום ב-#shorthebrewpapereviews:

BuboGPT: Enabling Visual Grounding in Multi-Modal LLMs

המאמר משלב כמה מודלים חזקים שאומנו למשימות הקשורות ל-grounding ויזואלי כמו Recognize Anything Model שמפיק את הקטגוריות של האובייקטים. לאחר מכן מזינים הקטגוריות אלו למודל Ground-DINO שמזהה את מיקום האובייקטים. בסוף מכניסים את התוצאה ל-SAM ועושים סגמנטציה עדינה של האובייקטים

בשלב האחרון צריך למצוא את הזוג (קטגוריה, מסכה) המתאים לשאילתה שלכם. בשביל כך מעבירים את השאילתה דרך GPT4 ומחפשים את הקטגוריה הקרובה ביותר מהרשימה שהוצאה על יד SAM. אוקיי סיימנו עם grounding ויזואלי אבל הבטחתי לכם 3 מודים - אז איפה נעלם האודיו?

קודם כל ״מיישרים״ את הייצוגי אודיו ותמונות על ידי אימון שני אנקודרים (לכל אחד) על דאטהסט של מכיל זוגות של אודיו ותמונה מתאימה. גם כאן מכיילים מודלים קיימים: BLIP2 לתמונה ו-ImageBind לאודיו. בסוף מכיילים את כל המפלצת הזו על שילובים של אודיו-טקסט, תמונה-טקסט, ותמונה-טקסט-אודיו.

https://huggingface.co/papers/2307.08581