# Review 121: [Short] AUDIOLDM 2: LEARNING HOLISTIC AUDIO GENERATION WITH SELF-SUPERVISED PRETRAINING, 11.08.23 https://huggingface.co/papers/2308.05734

**Paper: https://arxiv.org/abs/2308.05734v3**

רוצים לגנרט/לערוך אות אודיו ממגוון סוגים של דאטה? מתברר שניתן ליצור אות אודיו מתמונה, וידאו, טקסט ואפילו אות IMU (המתאר אופייני תנועה של אובייקט כמו מהירות כיוונית ותאוצה).

היום ב #shorthebrewpapereviews סוקרים מאמר שמציע מודל המסוגל לעשות את זה. בגדול במשימות מולטימודליות עיקר הבעיה היא למפות את כל סוגי הדאטה למרחב השיכון(embedding) באופן הגיוני. כלומר השיכון של אודיו של ציוץ ציפורים צריך להיות קרוב לשיכון של טקסט ״ציוץ של ציפורים״. המחברים משתמשים במודלים מאומנים עבור כל סוגי הדאטה שהם יוצרים אודיו מהם כדי למפות את הדאטה למרחב השיכון (למשל לוידאו ול-imu משתמשים ב-ImageBind, לטקסט ב-Flan-T5, ובאודיו ב-CLAP).

לאחר מכן הם מכיילים gpt-2 שמטרתו להעביר את כל הייצוגים לאותו מרחב השיכון. לאחר מכן לוקחים את הפלט של gpt-2 ומעבירים אותו למודל דיפוזיה לטנטי המאומן לגנרט אודיו בהתאם לתנאי (המקודד באמצעות וקטור השיכון). את מודל הדיפוזיה הלטנטי מאמנים על דאטה אודיו לא מתויג כאשר לפני כניסת האות לאנקודר מעבירים אותו לתדרי מל (mel frequencies).

לאחר מכן מכיילים את המודל עם הדאטה המתויג (אודיו והתיאור שלו). פרט נוסף מעניין לגבי המאמר הוא שבמהלך האימון ב 75% מהמקרים משתמשים בשיכון של התיאור כדי לגנרט אודיו וב- 25% הנותרים מכניסים במקומו את הייצוג של האודיו העצמו המופק באמצעות מודל AudioMAE (אנקודר של אותו אודיו).