# Review 161: [Short] PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis

**Paper: https://arxiv.org/abs/2310.00426v3**

https://arxiv.org/abs/2310.00426

מאמר חמוד שמציע זירוז תהליכי האימון של מודלי דיפוזיה שידועים כמאוד יקרים ויוצרים כמויות גדולות של פליטת CO2. בגדול המאמר מציע לשלב כמה גישות לאימון שהוצעו בזמן האחרון. אז היום ב-#shorthebrewpapereviews סוקרים מאמר עם השם שמכיל מילה ״אומנות״ (משך את תשומת ליבי ללא ספק והיווה אחת הסיבות לכך שעיני נתפסה על המאמר הזה).

אוקיי, אז מה הם בעצם עשו? קודם כל הם לקחו דאטהסט של תמונות (כמו LAION) ויצרו כותרות של התמונות בו באמצעות מודל חזק הנקרא LLAVA הטענה במאמר שכך נוצרות כותרות עשירות (סמנטית) הרבה יותר מהדאטהסט המקורי דבר שני, הם השתמשו בארכיטקטורת הטרנספורמרים כמודל לשערוך הרעש במודל דיפוזיה (במקום UNet).

כלומר הם לקחו מודל דיפוזיה שפועל בתחום לטנטי (stable diffusion) והחליפו UNet בהרבה שכבות של טרנספורמרים (זה הוצע במאמר Scalable Transformers למיטב ידיעתי) הם גם שכללו את שכבת הנרמול בטרנספורמר שאפשר להם לחתוך את כמות הפרמטרים בצורה משמעותית. בשלב האחרון הם כיילו את המודל באמצעות high-quality aesthetic data (הם הסבירו במאמר איך הם יוצרים אותו)

שילוב של כל הגישות האלו אפשר להקטין את זמן האימון (ופליטות הגז) בצורה משמעותית ד״א הם השתמשו בגרסאות המוקפאות של מודל T5 כדי לבנות ייצוג של טקסט וב- VAE מהמאמר על מודלי דיפוזיה לטנטיים לבניית ייצוג התמונה.