# Review 114: [Short] WOUAF: Weight Modulation for User Attribution and Fingerprinting in Text-to-Image Diffusion Models https://huggingface.co/papers/2306.04744

**Paper: https://arxiv.org/abs/2306.04744v3**

נניח שהצלחתם להנדס פרומפט מאוד מוצלח ל-MidJourney והוא צייר לכם תמונה מדהימה.אם אתם לא רוצים שאף אחד ישתמש בתמונה הזו בלי לתת לכם קרדיט. אבל איך אתם מוכיחים שאתם בעצם יצרתם את התמונה הזו? אתם צריכים להחביא בתמונה איזה סימן מים משלכם (watermark) כדי שתוכלו להוכיח את בעלותכם.

היום ב-#shorthebrewpapereviews סוקרים מאמר המציע שיטה להוספה של סימן מים המאפשרת להגיד האם תמונה נתונה נוצרה על ידיכם. יש שתי דרישות מסימני מים על התמונות המגונרטות. הדרישה הראשונה שהתמונות שנוצרו עבור אותו פרומפט (ואותו seed) עם ובלי סימן מים צריכות להיות מאוד דומות. הדרישה השנייה היא שניתן לשחזר את סימן המים הזה מהתמונה בצורה יחסית מדויקת. המחברים מציעים להוסיף את סימן המים הזה למפענח(decoder) של מודל Stable Diffusion.

איך זה נעשה? קודם כל מגרילים וקטור בינארי, מקודדים אותו עם רשת A מאומנת כדי ליצור מסכות לכל שכבה. למעשה כל נוירון של הפלט של כל שכבה של הדקודר מוכפל בפלט של A (כל שכבה ממוסכת בנפרד). כדי לשחזר סימן מים מאמנים עוד רשת שהקלט F (שהיא ResNet50) שלה היא תמונה והפלט שלה היא סימן המים עליה. פונקציית לוס כאן מורכבת משני איברים: הראשון מיועד לאימון של רשת F (לוס בינארי על כל ביט של סימן המים) והשני דואג שהסופת של סימן המים לא ישנה את התמונה יותר מדי דרך מזעור של perceptual loss בין התמונה עם סימן מים וזו שבלעדיו.