# Review 127: [Short] Dual-Stream Diffusion Net for Text-to-Video Generation, 17.08.23 https://huggingface.co/papers/2308.08316

**Paper: https://arxiv.org/abs/2308.08316v3**

הגישות הראשונות ליצירה של וידאו מטקסט באמצעות מודלי דיפוזיה יצאו לראשונה לפני שנה וחצי ומאז השתדרגו משמעותית מבחינת איכות הוידאו, אורכו והתאמתו לתיאור. היום ניתן לעשות גם פרסונליזציה למודלים אלו כלומר ליצור וידאו עם אובייקט ספציפי (החתול שלכם).

המאמר שנסקור היום ב-#shorthebrewpapereviews משדרג את הגישה הזו ומאפשר ליצור וידאו לא רק לאובייקט מסוים אלא גם לדפוס תנועה מסיום (הנגזר מוידאו אחר למשל). וכל זה בהתאמה לתיאור הטקסטואלי. איך מאמנים מודל כזה? מזינים למודל את הוידאו ובשלב הראשון מעבירים כל פריים דרך האנקודר להפקה של ייצוגו הלטנטי (עם VQ-VAE).

מהייצוג הזה מאמנים מודל המפרק את הייצוג הזה את ייצוג התנועה בוידאו (בין הפריימים) לבין ייצוג התוכן של הוידאו (כל אחד מהם הוא מערך של וקטורי ייצוג) – זה נעשה באמצעות Motion Decomposer. מערך וקטורים אלו מוזן למודל דיפוזיה משלו (מכאן בא השם dual stream) שעושים את קסמיהם הרגילים. פלטי מודלי דיפוזיה אלו מוזנים לרשת ש״מערבבת״ אותם ומוציאה שני ייצוגים מסונכרנים של תנועה ושל התוכן.

בסוף שני ייצוגים מסוכנרנים אלו מוזן לרשת המשלבת אותם ובונה ייצוג של וידאו שעובר דרך הדקודר כדי לגנרט וידאו. כאשר רוצים לגנרט וידאו לייצוג תנועה נתון מכיילים את המודל על ידי מזעור לוס השחזור את ייצוג התנועה מהייצוג הוידאו המגונרט.