

# Modele dyfuzyjne w zadaniu super-rozdzielczości wideo na konsumenckich kartach graficznych

Daniel Machniak<sup>1</sup>

<sup>1</sup> Politechnika Warszawska, plac Politechniki 1, 00-661 Warszawa, Polska

**Streszczenie.** Adaptacja modeli dyfuzyjnych typu Diffusion Transformer (DiT) do zadania super-rozdzielczości wideo (VSR) wiąże się z wysokimi wymaganiami pamięciowymi, często uniemożliwiającymi inferencję na sprzęcie konsumenckim. Niniejszy artykuł prezentuje zoptymalizowany potok przetwarzania oparty na architekturze FlashVSR, umożliwiający efektywne uruchomienie modelu na kartach graficznych z 10 GB VRAM. Wdrożono strategię kafelkowania przestrzennego oraz zastąpiono standardowe mechanizmy uwagi wydajnymi wariantami: kwantyzowanym SageAttention oraz dynamicznie rzadkim SpargeAttention. Wyniki eksperymentów na zbiorach REDS i VideoLQ potwierdzają, że proponowane podejście skutecznie redukuje narzut pamięciowy przy marginalnym spadku wierności rekonstrukcji, czyniąc zaawansowane metody VSR bardziej dostępnymi.

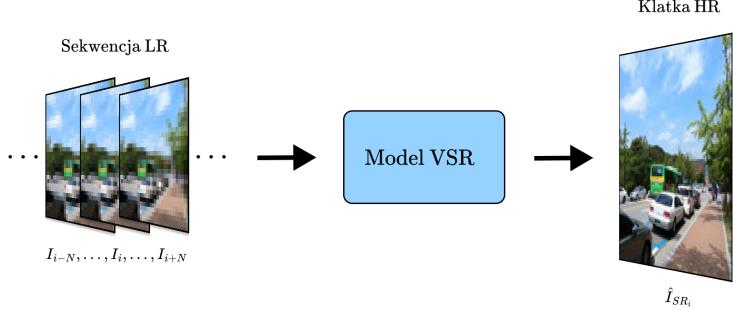
**Słowa kluczowe:** Super-rozdzielcość wideo · Modele dyfuzyjne · FlashVSR · Transformery wizyjne · Optymalizacja pamięciowa

## 1 Wprowadzenie

Super-rozdzielcość wideo (ang. *Video Super-Resolution*, VSR) stanowi kluczowe zagadnienie w dziedzinie niskopoziomowego widzenia komputerowego, którego celem jest rekonstrukcja sekwencji wideo o wysokiej rozdzielczości (HR) z materiałów wejściowych o niskiej rozdzielczości (LR) [1]. W przeciwieństwie do super-rozdzielczości pojedynczego obrazu (SISR), zadanie to wymaga efektywnego wykorzystania korelacji czasowych oraz informacji zawartych w sąsiednich klatkach w celu odzyskania brakujących detali i zachowania spójności czasowej [1].

Dynamiczny rozwój głębokiego uczenia (ang. Deep Learning) w ostatniej dekadzie doprowadził do powstania zaawansowanych architektur, początkowo opartych na konwolucyjnych sieciach neuronowych (CNN), a następnie na Transformerach, które zdominowały zadania przetwarzania sekwencji [1, 2].

W ostatnich latach szczególną uwagę badaczy przyciągają probabilistyczne modele dyfuzyjne (ang. *Denoising Diffusion Probabilistic Models*, DDPM), które dzięki iteracyjnemu procesowi odszumiania pozwalają na generowanie próbek o jakości przewyższającej tradycyjne podejścia, takie jak GAN czy VAE [3]. Ewolucja tych systemów doprowadziła do powstania architektury Diffusion Transformer (DiT), która zastępuje tradycyjny szkielet U-Net mechanizmem uwagi, oferując lepszą skalowalność i efektywność w zadaniach generatywnych [4]. Jednakże adaptacja modeli DiT do zadania VSR wiąże się z istotnymi wyzwaniami oblicze-



**Rysunek 1.** Ogólny schemat procesu super-rozdzielczości wideo (VSR). Model rekonstruuje klatkę wysokiej rozdzielczości na podstawie sekwencji klatek wejściowych.

niami. Mechanizm uwagi charakteryzuje się kwadratową złożonością czasową i pamięciową  $O(N^2)$  względem długości sekwencji wejściowej. W kontekście wideo, gdzie sekwencja tokenów obejmuje wymiary przestrzenne i czasowe, prowadzi to do zaporowego zapotrzebowania na pamięć GPU, co często uniemożliwia inferencję na sprzątce konsumenckim.

Niniejszy artykuł podejmuje problem optymalizacji dyfuzyjnych modeli VSR w celu ich efektywnego uruchomienia na kartach graficznych o ograniczonej pamięci VRAM. Głównym celem pracy jest implementacja potoku przetwarzania opartego na architekturze FlashVSR, zintegrowanego z technikami redukcji narzutu pamięciowego. W szczególności badano zastosowanie kafelkowania przestrzenno-czasowego oraz nowoczesnych algorytmów uwagi, takich jak FlashAttention [5] oraz kwantyzowane SageAttention [6]. Podejście to ma na celu przełamanie bariery sprzętowej przy zachowaniu wysokiej wierności rekonstrukcji obrazu.

## 2 Tło teoretyczne i przegląd literatury

Historycznie metody VSR ewoluowały od prostych algorytmów interpolacyjnych do złożonych systemów sztucznej inteligencji. Wczesne podejścia oparte na głębokim uczeniu wykorzystywały dwuwymiarowe splotowe sieci neuronowe (2D CNN), które traktowały wideo jako zbiór niezależnych obrazów lub wykorzystywały proste mechanizmy fuzji czasowej. Przełomem okazało się wprowadzenie dedykowanych modułów do wyrównywania klatek, takich jak deformowalne sploty (Deformable Convolution) zastosowane w modelu EDVR, czy mechanizmy propagacji rekurencyjnej w BasicVSR. Rozwiązania te, choć skuteczne, często borykają się z ograniczeniami w modelowaniu długodystansowych zależności czasowych [7].

Równolegle, sukces architektury Transformer w przetwarzaniu języka naturalnego zainspirował jej adaptację do zadań wizyjnych. Transformer wizyjny (ang. *Vision Transformer*, ViT) [2] zastąpił lokalne operacje splotowe globalnym

mechanizmem uwagi, co pozwoliło na lepsze uchwycenie kontekstu globalnego obrazu. W kontekście generatywnym, modele dyfuzyjne (DDPM) zdeterminizowały sieci GAN, oferując stabilniejszy trening i wyższą jakość generowanych próbek. Połączenie tych dwóch nurtów doprowadziło do powstania architektury Diffusion Transformer (DiT) [4], która skaluje się efektywnie niż modele dyfuzyjne oparte na sieci U-Net. Niniejsza praca osadzona jest w tym najnowszym nurcie, adaptując architekturę DiT do zadania VSR przy jednoczesnym rozwiązaniu problemów wydajnościowych.

## 2.1 Sformułowanie problemu VSR

Problem super-rozdzielczości wideo jest zdefiniowany jako zadanie odwrotne, w którym dążymy do odzyskania sekwencji wysokiej rozdzielczości (HR) na podstawie obserwowanej sekwencji o niskiej rozdzielczości (LR). Proces powstawania materiału LR jest zazwyczaj modelowany jako złożenie degradacji fizycznych i cyfrowych.

Ogólny model degradacji dla i-tej klatki wideo można zapisać jako funkcję zależną od klatki docelowej  $\hat{I}_i$  oraz jej sąsiedztwa czasowego [7]:

$$I_i = \phi\left(\hat{I}_i, \{\hat{I}_j\}_{j=i-N}^{i+N}; \theta_\alpha\right), \quad (1)$$

gdzie  $I_i$  oznacza obserwowaną klatkę o niskiej rozdzielczości,  $\hat{I}_i$  to sekwencja oryginalna o wysokiej rozdzielczości,  $N$  oznacza promień czasowy (zakres sąsiednich klatek), a  $\theta_\alpha$  reprezentuje parametry procesu degradacji (np. szum, rozmycie).

W bardziej szczegółowym ujęciu, uwzględniającym ruch między klatkami oraz standardowe operacje przetwarzania sygnału, proces degradacji dla sąsiadnej klatki  $j$  względem klatki referencyjnej  $i$  jest definiowany jako [7]:

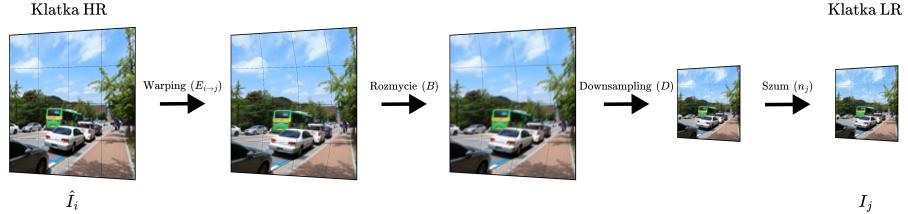
$$I_j = DBE_{i \rightarrow j}\hat{I}_i + n_j, \quad (2)$$

gdzie  $D$  oznacza operator podpróbkowania (downsampling),  $B$  reprezentuje operator rozmycia (blur), a  $n_j$  to addytywny szum. Kluczowym elementem w kontekście wideo jest operator  $E_{i \rightarrow j}$ , który oznacza operację zniekształcenia (warping) zgodną z ruchem od klatki  $i$  do  $j$ .

Model ten zakłada, że klatka LR powstaje poprzez przekształcenie geometryczne klatki HR, jej rozmycie, zmniejszenie rozdzielczości oraz dodanie szumu.

Celem modelu VSR jest znalezienie funkcji odwzorowującej  $f_{VSR}$ , sparametryzowanej przez wagę  $\theta_{f_{VSR}}$ , która estymuje klatkę wysokiej rozdzielczości  $\hat{I}_{SR_i}$  na podstawie sekwencji klatek wejściowych LR [7]:

$$\hat{I}_{SR_i} = f_{VSR}\left(I_i, \{I_j\}_{j=i-N}^{i+N}; \theta_{f_{VSR}}\right), \quad (3)$$

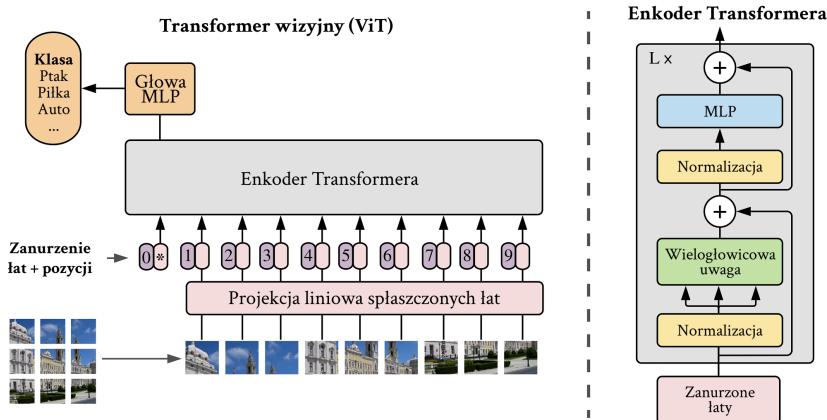


**Rysunek 2.** Ilustracja modelu degradacji. Klatka HR podlega przekształceniom geometrycznym, rozmyciu, podpróbkowaniu i zaszumieniu, tworząc klatkę LR.

## 2.2 Transformer wizyjny (ViT)

Architektura Transformer, która stała się standardem w przetwarzaniu języka naturalnego, znalazła skuteczne zastosowanie w wizji komputerowej pod postacią Tranformerera wizyjnego (ang. *Vision Transformer*, ViT). W przeciwieństwie do splotowych sieci neuronowych (CNN), które polegają na lokalnych operacjach splotu i wbudowanych założeniach indukcyjnych dotyczących lokalności i niezmienniczości przesunięcia, ViT interpretuje obraz jako sekwencję lat (ang. patches), przetwarzając je za pomocą standardowego enkodera Transformer [2].

W modelu ViT obraz wejściowy  $x \in \mathbb{R}^{H \times W \times C}$  jest dzielony na sekwencję spłaszczonych lat 2D  $x \in \mathbb{R}^{N \times (P^2 \cdot C)}$ , gdzie  $(P, P)$  to rozdzielcość pojedynczej łaty, a  $N = HW/P^2$  stanowi efektywną długość sekwencji wejściowej. Każda łata jest następnie rzutowana liniowo do stałego wymiaru ukrytego  $D$ , a do uzyskanych



**Rysunek 3.** Schemat działania Transformera wizyjnego (ViT). Obraz dzielony jest na łaty, rzutowany liniowo i przetwarzany przez warstwy atencji.

wektorów dodawane są wyuczalne zanurzenia pozycyjne, aby zachować informacje o strukturze przestrzennej obrazu. Tak przygotowana sekwencja tokenów jest przetwarzana przez warstwy wieloglówkowej uwagi (ang. *Multi-Head Self-Attention*, MSA), co pozwala modelowi na integrację informacji z całego obrazu już w pierwszych warstwach sieci, w przeciwieństwie do ograniczonego pola recepcyjnego w CNN [2].

Bezpośrednia adaptacja mechanizmu MSA do materiałów wideo wiąże się z opisany we wstępnie problemem eksplozji liczby tokenów  $N$ . Ponieważ sekwencja obejmuje wymiar czasowy, standardowa macierz uwagi staje się wąskim gardłem, co motywuje poszukiwanie wariantów atencji o zredukowanej złożoności lub zastosowanie technik okienkowych.

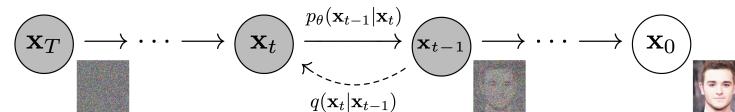
### 2.3 Modele dyfuzyjne i architektura Diffusion Transformer (DiT)

Probabilistyczne modele dyfuzyjne odszumiania (DDPM) zrewolucjonizowały dziedzinę syntez obrazów, oferując wyższą jakość generowanych próbek i większą różnorodność w porównaniu do wcześniejszych architektur GAN [4]. Działanie tych modeli opiera się na dwóch procesach: ustalonym procesie „w przód”, który stopniowo dodaje szum Gaussa do danych aż do uzyskania czystego szumu, oraz wyuczalnym procesie „wstecz”, który iteracyjnie odtwarza strukturę danych z szumu, modelując rozkład warunkowy  $p_\theta(x_{t-1} | x_t)$  [3].

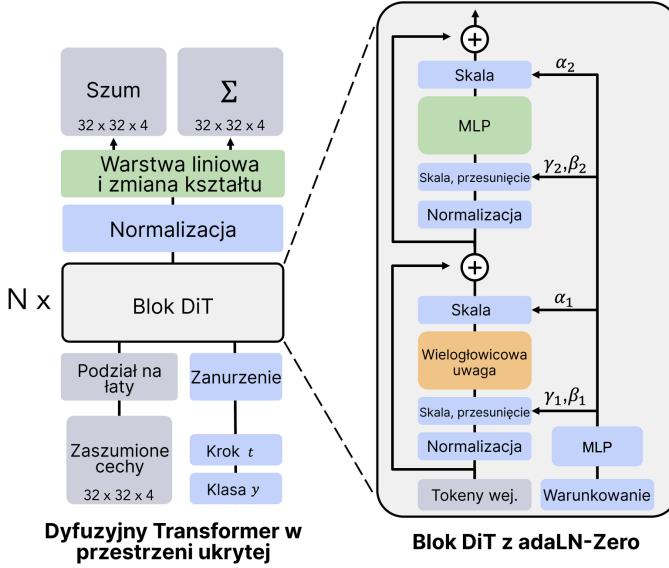
Tradycyjnie, jako szkielet (ang. *backbone*) dla procesu odszumiania wykorzystywano architektury U-Net opartą na splotach.

Ostatnie badania, w tym praca Peeblesa i Xie [4], zaproponowały nową klasę modeli określającą jako Diffusion Transformers (DiT), która zastępuje tradycyjny U-Net architekturą Transformera działającą na reprezentacji utajonej (ang. *latent space*). W podejściu tym obraz wejściowy zakodowany przez autoenkoder wariacyjny (ang. *variational autoencoder*, VAE) jest dzielony na sekwencję lat, analogicznie jak w modelu ViT, a następnie przetwarzany przez standardowe bloki transformera.

Kluczowym elementem adaptacji Transformera do zadań generatywnych jest mechanizm warunkowania. W architekturze DiT zastosowano warstwę Adaptive Layer Normalization (adaLN), która wykorzystuje parametry skali i przesunięcia w warstwach normalizacyjnych na podstawie wektorów osadzenia czasu (ang. *timestep*) i warunku (np. etykiety klasy lub obrazu LR).



**Rysunek 4.** Graf probabilistyczny modelu dyfuzyjnego. Proces w przód ( $q$ ) degraduje obraz do szumu, a proces wsteczny ( $p_\theta$ ) rekonstruuje obraz.



**Rysunek 5.** Architektura DiT (po lewej) oraz szczegółowa budowa bloku DiT z mechanizmem adaLN-Zero (po prawej), sterującym procesem generacji na podstawie kroku czasowego  $t$  i warunku  $y$ .

Główną zaletą architektury DiT jest jej przewidywalna skalowalność. Autorzy wykazali silną korelację między złożonością obliczeniową modelu, a jakością generowanych obrazów - zwiększenie głębokości lub szerokości sieci prowadzi do systematycznej poprawy wyników, co czyni DiT atrakcyjnym wyborem dla zadań wymagających wysokiej wierności, takich jak VSR.

### 3 Proponowana metoda i optymalizacja

W niniejszym rozdziale przedstawiono proponowane podejście do optymalizacji inferencji modeli super-rozdzielczości wideo na sprzęcie o ograniczonych zasobach pamięciowych. Opracowany potok przetwarzania integruje model FlashVSR z implementacją mechanizmów zarządzania pamięcią oraz zoptymalizowanymi jądrami obliczeniowymi atencji. Głównym założeniem jest redukcja narzutu pamięci VRAM poprzez dekompozycję problemu w wymiarze przestrzennym (kafelkowanie) oraz redukcję precyzji obliczeń w warstwach atencji, przy jednoczesnym zachowaniu spójności generowanej struktury obrazu.

#### 3.1 Bazowa architektura FlashVSR

Jako fundament proponowanego rozwiązania przyjęto architekturę FlashVSR [8], która stanowi nowatorskie podejście do zagadnienia super-rozdzielczości wideo w

trybie strumieniowym. W przeciwnieństwie do standardowych modeli dyfuzyjnych, wymagających kosztownego obliczeniowo i wieloetapowego procesu odszumiania, FlashVSR wykorzystuje zaawansowany, trójetapowy proces destylacji wiedzy. Pozwala to na inferencję w pojedynczym kroku, w którym model ucznia mapuje szum w przestrzeni ukrytej bezpośrednio do czystego obrazu, co jest kluczowe dla zastosowań w czasie rzeczywistym.

Istotnym elementem architektury, zapewniającym efektywność w przetwarzaniu sekwencji wideo, jest adaptacja mechanizmu atencji do przetwarzania przyczynowego (ang. *causal processing*). W tym podejściu generacja bieżącej klatki zależy wyłącznie od informacji zawartych w klatkach poprzednich, co umożliwia zastosowanie mechanizmu buforowania kluczy i wartości (KV Cache), techniki zapożyczonej z dużych modeli językowych. Dzięki temu cechy wyekstrahowane z poprzednich kroków czasowych są przechowywane w pamięci VRAM, eliminując konieczność ich ponownego obliczania i drastycznie redukując liczbę operacji w potoku przetwarzania wideo.

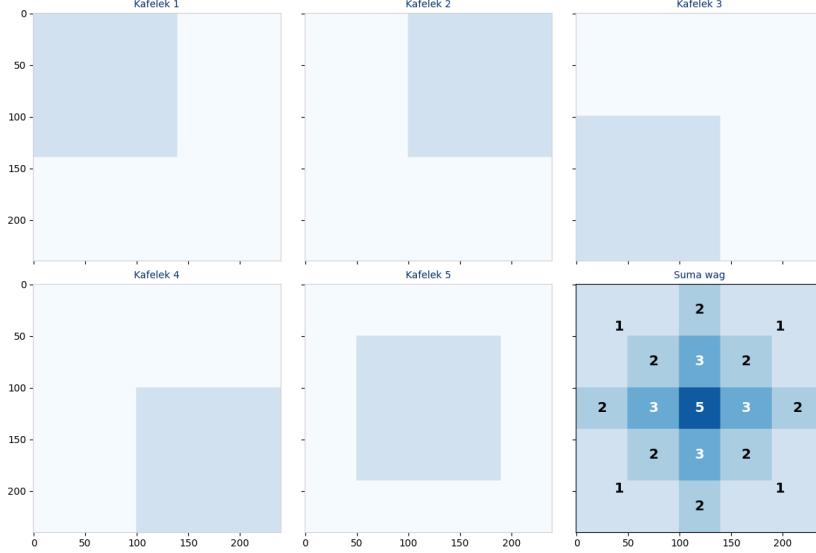
Kolejnym wyróżnikiem modelu jest zastosowanie lokalnie ograniczonej rzadkiej atencji (ang. *Locality-Constrained Sparse Attention*). Standardowe mechanizmy atencji globalnej często wykazują trudności z generalizacją do rozdzielczości wyższych niż te wykorzystane w procesie treningowym, co objawia się artefaktami w postaci powtarzających się wzorców geometrycznych. FlashVSR rozwiązuje ten problem poprzez nałożenie maski rzadkości, która ogranicza pole recepcyjne każdego tokenu do lokalnego sąsiedztwa przestrzenno-czasowego. Podejście to nie tylko poprawia jakość rekonstrukcji na ultra-wysokich rozdzielczościach poprzez eliminację błędów pozycyjnych, ale także znaczaco zmniejsza złożoność obliczeniową operacji mnożenia macierzy.

Uzupełnieniem architektury jest zoptymalizowany moduł dekodujący, określany jako Tiny Conditional Decoder. W tradycyjnych architekturach opartych na dyfuzji, dekoder wariacyjnego autoenkodera często stanowi wąskie gardło wydajnościowe. FlashVSR zastępuje go lekkim wariantem warunkowym, który oprócz reprezentacji utajonej przyjmuje jako wejście również przeskalowaną klatkę niskiej rozdzielczości. Wykorzystanie informacji strukturalnych bezpośrednio z obrazu wejściowego jako sygnału pomocniczego pozwala na znaczne odciążenie sieci dekodującej i redukcję jej głębokości, przy jednoczesnym zachowaniu wysokiej wierności detali.

### 3.2 Strategia kafelkowania przestrzenno-czasowego

W celu przezwyciężenia ograniczeń pamięci VRAM, uniemożliwiających przetwarzanie całego wideo wysokiej rozdzielczości, zastosowano dekompozycję danych. W domenie czasowej dłuża sekwencja dzielona jest na nakładające się klipy, które są przetwarzane niezależnie, a ich spójność na granicach zapewnia uśrednianie predykcji.

Kluczowym elementem implementacji jest kafelkowanie przestrzenne. Wideo wejściowe jest dzielone na regularną siatkę, nakładających się na siebie, fragmentów o ustalonej rozdzielczości (w eksperymentach przyjęto  $192 \times 192$  pikseli). Tak



**Rysunek 6.** Wizualizacja strategii kafelkowania. Obraz jest dzielony na nakładające się fragmenty, a wagi (prawy dolny róg) zapewniają płynne przejścia na granicach kafelków.

zdefiniowane fragmenty są przetwarzane przez model sekwencyjnie, co pozwala na utrzymanie stałego, niskiego zużycia pamięci niezależnie od rozdzielczości materiału wejściowego. Rekonstrukcja pełnej klatki polega na złożeniu przetworzonych fragmentów w jedną całość, przy czym na granicach kafelków ostateczna wartość pikseli jest obliczana poprzez uśrednienie predykcji. Podejście to skutecznie eliminuje widoczność szwów łączenia, zapewniając spójność strukturalną obrazu wynikowego przy minimalnym narzucie obliczeniowym.

### 3.3 Optymalizacja mechanizmu uwagi

Mimo zastosowania technik kafelkowania, standardowa implementacja mechanizmu uwagi nadal stanowi istotne obciążenie dla przepustowości pamięci i jednostek obliczeniowych konsumenckich kart graficznych. W oryginalnej architekturze FlashVSR wykorzystano mechanizm lokalnie ograniczonej rzadkiej uwagi, który narzuca statyczne okno przetwarzania w celu eliminacji błędów generalizacji pozycyjnej. W niniejszej pracy zmodyfikowano to podejście, zastępując bazowe jądra obliczeniowe rozwiązaniami bardziej efektywnymi pamięciowo i obliczeniowo. Zamiast standardowego algorytmu FlashAttention [5] wdrożono SageAttention [6], natomiast statyczną rzadkość blokową zastąpiono dynamicznym mechanizmem SpargeAttention [9].

W celu zminimalizowania kosztownych transferów danych między pamięcią główną HBM a rdzeniami obliczeniowymi, zastosowano metodę SageAttention

[6], która umożliwia efektywne wykonywanie operacji w precyzyji 8-bitowej. Głównym wyzwaniem w kwantyzacji mechanizmu uwagi jest występowanie wartości odstających (ang. *outliers*) w kanałach macierzy kluczy ( $K$ ), co w naiwnych implementacjach prowadzi do znacznej degradacji jakości generowanego obrazu, objawiającej się rozmyciem tekstur. Zaimplementowane rozwiążanie adresuje ten problem poprzez technikę wygładzania macierzy  $K$  (ang. *K-smoothing*), polegającą na odjęciu średniej wartości kanału od macierzy przed kwantyzacją. Operacja ta centruje rozkład wartości, eliminując dominujące odchylenia bez wpływu na wynik końcowy funkcji Softmax, co pozwala na zachowanie wysokiej precyzyji operacji mnożenia macierzy  $Q$  i  $K$  w zredukowanej reprezentacji bitowej.

Uzupełnieniem kwantyzacji jest redukcja liczby operacji realizowana przez mechanizm SparseAttention [9], który wprowadza dynamiczną selekcję bloków w czasie rzeczywistym bez konieczności dodatkowego treningu. Proces ten przebiega dwuetapowo. W pierwszej fazie następuje szybka predykcja istotności: bloki macierzy zapytań ( $Q$ ) i kluczy ( $K$ ) są kompresowane do wektorów reprezentujących ich wartości średnie, co pozwala na efektywne obliczenie estymowanego podobieństwa. Jeżeli wynik tej operacji dla danej pary bloków znajduje się poniżej ustalonego progu, odpowiadający im fragment macierzy uwagi jest klasyfikowany jako nieistotny i nie jest pobierany z pamięci, co eliminuje transfer danych. W drugiej fazie, realizowanej bezpośrednio na poziomie wątków GPU podczas obliczania funkcji Softmax, weryfikowany jest wkład danego bloku w globalną sumę normalizacyjną. W przypadku gdy wkład ten jest znikomy, algorytm dynamicznie rezygnuje z operacji mnożenia przez macierz wartości ( $V$ ), wykonując obliczenia wyłącznie dla elementów determinujących wynik końcowy.

## 4 Eksperymenty i analiza wyników

W celu zweryfikowania skuteczności proponowanego potoku przetwarzania, przeprowadzono eksperymenty porównawcze, mające na celu zbadanie wpływu technik optymalizacyjnych na jakość generowanego obrazu. Ewaluację oparto na uznanych w środowisku naukowym zbiorach danych oraz zestawie zróżnicowanych metryk jakościowych.

### 4.1 Zbiory danych

Do weryfikacji wyników wykorzystano dwa standardowe benchmarki o odmiennej charakterystyce: zbiór REDS [10], zawierający wysokiej jakości sekwencje o dużej dynamice ruchu, wykorzystywany do oceny wierności rekonstrukcji względem idealnego wzorca, oraz zbiór VideoLQ [11], składający się z rzeczywistych nagrań o niskiej jakości pobranych z serwisów internetowych, który posłużył do sprawdzenia zdolności generalizacji modelu w obecności złożonych, niesyntetycznych degradacji.

## 4.2 Metryki oceny

Ocena jakości rekonstrukcji została przeprowadzona uwzględniając zarówno wierność sygnału względem oryginału, jak i subiektywną percepcję wizualną. W grupie metryk referencyjnych zastosowano klasyczny wskaźnik PSNR [12] (Peak Signal-to-Noise Ratio), który mierzy błąd średniokwadratowy między pikselami obrazu rekonstruowanego a referencyjnego. Mimo powszechności stosowania, PSNR jest krytykowany za słabą korelację z ludzkim postrzeganiem jakości, często faworyzując obrazy gładkie kosztem utraty detali teksturalnych. Dlatego też wyniki uzupełniono o wskaźnik SSIM [13] (Structural Similarity Index), który lepiej oddaje zmiany w strukturze obrazu, luminancji i kontraste. Najbardziej zaawansowaną miarą w tej grupie jest LPIPS [14] (Learned Perceptual Image Patch Similarity), obliczający dystans percepcyjny w przestrzeni cech głębokiej sieci neuronowej, co pozwala na znacznie precyzyjniejszą ocenę zgodności tekstur i struktur niż proste operacje na pikselach.

W przypadku braku idealnego wzorca oraz w celu oceny estetyki generowanych obrazów, posłużono się metrykami bezreferencyjnymi. Wykorzystano wskaźnik NIQE [15] (Naturalness Image Quality Evaluator), badający odchylenia statystyk obrazu od modelu naturalnych scen, co pozwala na wykrycie nienaturalnych artefaktów generacji. Zastosowano również nowoczesne metody oparte na głębokim uczeniu, takie jak MUSIQ [16] (Multi-scale Image Quality Transformer), który dzięki architekturze Transformer ocenia jakość na wielu skalach jednocześnie, oraz CLIPQA [17], wykorzystującą model językowo-wizualny CLIP do oceny semantycznej zgodności obrazu z pozytywnymi wzorcami estetycznymi. Całość uzupełnia specjalistyczna metryka wideo DOVER [18] (Disentangled Objective Video Quality Evaluator), która dekomponuje ocenę na dwa niezależne aspekty: jakość techniczną, wrażliwą na szумy i rozmycia, oraz jakość estetyczną, związaną z kompozycją i stylem, co jest kluczowe dla pełnej oceny modeli generatywnych.

## 4.3 Środowisko testowe

Eksperymenty przeprowadzono na stacji roboczej wyposażonej w układ graficzny NVIDIA GeForce RTX 3080 z 10 GB pamięci VRAM. Rozmiar pojedynczego kafelka ustalono na  $192 \times 192$  pikseli, co stanowi wartość pozwalającą na stabilną inferencję. W celu zachowania spójności na granicach fragmentów i eliminacji artefaktów brzegowych, wprowadzono zakładkę o szerokości 24 pikseli.

## 4.4 Wyniki eksperymentów

Zestawienie wyników dla modelu referencyjnego, wersji z kafelkowaniem oraz pełnego potoku zoptymalizowanego przedstawiono w Tabeli 1. Analiza danych wskazuje, że dekompozycja obrazu oraz redukcja precyzji atencji wiążą się z marginalnym spadkiem parametrów jakościowych. Zastosowanie kafelkowania skutkuje obniżeniem wskaźnika PSNR na zbiorze REDS o 0.15 dB, co wynika z ograniczenia globalnego pola recepcyjnego do obszaru kafelka. Integracja zmodyfikowanych mechanizmów SageAttention oraz SparseAttention wpływa na wynik

**Tabela 1.** Porównanie jakości rekonstrukcji dla trzech badanych konfiguracji

Zbiór danych	Metryka	FlashVSR	FlashVSR + kafelkowanie	FlashVSR + kafelkowanie + modyfikacja uwagi
<b>REDS</b>	PSNR↑	23.31	23.16	23.13
	SSIM↑	0.6110	0.6075	0.6068
	LPIPS↓	0.3866	0.3950	0.3962
	NIQE↓	3.489	3.580	3.595
	MUSIQ↑	66.63	65.20	65.05
	CLIPQA↑	0.5221	0.5160	0.5152
<b>VideoLQ</b>	DOVER↑	12.66	12.15	12.08
	NIQE↓	4.070	4.150	4.165
	MUSIQ↑	52.27	51.40	51.28
	CLIPQA↑	0.3601	0.3540	0.3532
	DOVER↑	7.481	7.250	7.210

w stopniu pomijalnym, obniżając PSNR o kolejne 0.03 dB przy zachowaniu wysokiej wierności strukturalnej (SSIM).

## 5 Podsumowanie

Niniejsza praca podjęła problem barier sprzętowych ograniczających powszechnie zastosowanie dyfuzyjnych modeli Transformer w zadaniu super-rozdzielczości wideo. Zaproponowane rozwiązanie, integrujące architekturę FlashVSR z mechanizmem kafelkowania przestrzenno-czasowego oraz zoptymalizowanymi algorytmami uwagi (SageAttention, SpargeAttention), pozwoliło na skuteczne przeprowadzenie inferencji na konsumenckiej karcie graficznej z 10 GB pamięci VRAM. Uzyskane wyniki eksperymentalne jednoznacznie wskazują, że redukcja precyzji obliczeń oraz dekompozycja obrazu nie muszą wiązać się z widocznym pogorszeniem jakości generowanych treści. Obserwowany spadek wierności rekonstrukcji stanowi w pełni akceptowalny kompromis inżynierski w zamian za redukcję wymagań pamięciowych. Przedstawione podejście udowadnia, że generowanie wideo o wysokiej wierności jest osiągalne na szeroko dostępnym sprzęcie komputerowym.

## Bibliografia

1. Baniya, A.A., Lee, T.-K., Eklund, P.W., Aryal, S.: A Survey of Deep Learning Video Super-Resolution. IEEE Transactions on Emerging Topics in Computational Intelligence. 8, 2655–2676 (2024). <https://doi.org/10.1109/TETCI.2024.3398015>.

2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, <https://arxiv.org/abs/2010.11929>.
3. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models, <https://arxiv.org/abs/2006.11239>.
4. Peebles, W., Xie, S.: Scalable Diffusion Models with Transformers, <https://arxiv.org/abs/2212.09748>.
5. Dao, T., Fu, D.Y., Ermon, S., Rudra, A., Ré, C.: FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness, <https://arxiv.org/abs/2205.14135>.
6. Zhang, J., Wei, J., Huang, H., Zhang, P., Zhu, J., Chen, J.: SageAttention: Accurate 8-Bit Attention for Plug-and-play Inference Acceleration, <https://arxiv.org/abs/2410.02367>.
7. Liu, H., Ruan, Z., Zhao, P., Dong, C., Shang, F., Liu, Y., Yang, L., Timofte, R.: Video super-resolution based on deep learning: a comprehensive survey. Artificial Intelligence Review. 55, 5981–6035 (2022).
8. Zhuang, J., Guo, S., Cai, X., Li, X., Liu, Y., Yuan, C., Xue, T.: FlashVSR: Towards Real-Time Diffusion-Based Streaming Video Super-Resolution, <https://arxiv.org/abs/2510.12747>.
9. Zhang, J., Xiang, C., Huang, H., Wei, J., Xi, H., Zhu, J., Chen, J.: SpargeAttention: Accurate and Training-free Sparse Attention Accelerating Any Model Inference, <https://arxiv.org/abs/2502.18137>.
10. Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., Mu Lee, K.: NTIRE 2019 Challenge on Video Deblurring and Super-Resolution: Dataset and Study. W: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2019).
11. Chan, K.C.K., Zhou, S., Xu, X., Loy, C.C.: Investigating Tradeoffs in Real-World Video Super-Resolution, <https://arxiv.org/abs/2111.12704>.
12. Fardo, F.A., Conforto, V.H., Oliveira, F.C. de, Rodrigues, P.S.: A Formal Evaluation of PSNR as Quality Measurement Parameter for Image Segmentation Algorithms, <https://arxiv.org/abs/1605.07116>.
13. Nilsson, J., Akenine-Möller, T.: Understanding SSIM, <https://arxiv.org/abs/2006.13846>.
14. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, <https://arxiv.org/abs/1801.03924>.
15. Zvezdakova, A., Kulikov, D., Kondranin, D., Vatolin, D.: Barriers towards no-reference metrics application to compressed video quality analysis: on the example of no-reference metric NIQE, <https://arxiv.org/abs/1907.03842>.
16. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: MUSIQ: Multi-scale Image Quality Transformer, <https://arxiv.org/abs/2108.05997>.
17. Wang, J., Chan, K.C.K., Loy, C.C.: Exploring CLIP for Assessing the Look and Feel of Images, <https://arxiv.org/abs/2207.12396>.
18. Wu, H., Zhang, E., Liao, L., Chen, C., Hou, J., Wang, A., Sun, W., Yan, Q., Lin, W.: Exploring Video Quality Assessment on User Generated Contents from Aesthetic and Technical Perspectives, <https://arxiv.org/abs/2211.04894>.