VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

TVORBA REKLAMNÍHO VIDEA POMOCÍ NEURONOVÝCH MODELŮ

CREATING ADVERTISEMENT VIDEO USING NEURAL MODELS

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **EVGENIYA TAIPOVA**
AUTHOR

VEDOUCÍ PRÁCE **doc. RNDr. PAVEL SMRŽ, Ph.D.**
SUPERVISOR

BRNO 2024

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Zadání bakalářské práce

153488

Ústav:	Ústav počítačové grafiky a multimédií (UPGM)
Studentka:	**Taipova Evgeniya**
Program:	Informační technologie
Název:	**Tvorba reklamního videa pomocí neuronových modelů**
Kategorie:	Počítačová grafika
Akademický rok:	2023/24

Zadání:

1. Seznamte se s pokročilými neuronovými modely pro generování a editaci videa na základě textové interakce.
2. Shromážděte data pro průběžné vyhodnocování vytvářeného systému a přehled dostupných předtrénovaných modelů, které je možné použít.
3. Na základě získaných poznatků navrhněte a implementujte systém, který dokáže na základě zadaného reklamního sdělení a základního scénáře vytvořit video obsah.
4. Vyhodnoťte výsledky systému na shromážděných datech i v uživatelské studii, zaměřené na množství ušetřené práce při vytváření video obsahu.
5. Vytvořte stručný plakát prezentující práci, její cíle a výsledky.

Literatura:

- dle doporučení vedoucího

Při obhajobě semestrální části projektu je požadováno:

- funkční prototyp řešení

Podrobné závazné pokyny pro vypracování práce viz https://www.fit.vut.cz/study/theses/

Vedoucí práce:	**Smrž Pavel, doc. RNDr., Ph.D.**
Vedoucí ústavu:	Černocký Jan, prof. Dr. Ing.
Datum zadání:	1.11.2023
Termín pro odevzdání:	9.5.2024
Datum schválení:	21.12.2023

Abstrakt

Cílem této práce je vytvoření systému pro automatickou generaci reklamních videí z textových popisů, který usnadní uživatelům bez zkušeností ve videoprodukci ušetřit čas a peníze. Práce se skládá ze dvou hlavních částí. První část využívá generativní modely Stable Diffusion a Stable Video Diffusion pro tvorbu vizuálního obsahu a GPT-3.5 Turbo pro vytváření scénářů k reklamním videím. Druhá část je webová aplikace, která slouží uživatelům k zadávání potřebných informací pro reklamy a k zobrazení hotových videí. Tento systém zjednodušuje a urychluje proces tvorby různých typů reklam.

Abstract

The aim of this work is to create a system for the automatic generation of advertising videos based on textual descriptions, which will help users without video production experience save time and money. The work consists of two main parts. The first part uses generative models Stable Diffusion and Stable Video Diffusion for the creation of visual content and GPT-3.5 Turbo for creating scripts for advertising videos. The second part is a web application that allows users to input the necessary information for advertisements and to display the finished videos. This system simplifies and accelerates the process of creating various types of advertisements.

Klíčová slova

difúzní modely, reklama, neuronové modely, Stable Video Diffusion, umělá inteligence, generování videa, marketingové nástroje

Keywords

diffusion models, advertisement, neural models, Stable Video Diffusion, artificial intelligence, video generation, marketing tools

Citace

TAIPOVA, Evgeniya. *Tvorba reklamního videa pomocí neuronových modelů.* Brno, 2024. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce doc. RNDr. Pavel Smrž, Ph.D.

Tvorba reklamního videa pomocí neuronových modelů

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně pod vedením pana doc. RNDr. Pavela Smrža Ph.D. Uvedla jsem všechny literární prameny, publikace a další zdroje, ze kterých jsem čerpala.

.....
Evgeniya Taipova
8. května 2024

Poděkování

Ráda bych poděkovala doc. RNDr. Pavlu Smržovi, Ph.D. za vedení práce, cenné rady a věcné připomínky, které mi pomohly při tvorbě této práce. Také bych chtěla poděkovat své rodině a příteli za jejich podporu.

Obsah

1 Úvod **5**

2 Základy reklamního videa **6**
2.1 Typy reklamy 6
2.2 Cílová skupina a jedinečná prodejní nabídka 7

3 Modely převodu textu na video **8**
3.1 Převod textu na obraz jako základ pro převod textu na video 8
3.2 Vývoj video technologií 14

4 Přehled aktuálních nástrojů pro tvorbu videoreklamy **20**
4.1 Synthesia.io 20
4.2 Pictory.ai 20
4.3 Designs.ai 21
4.4 Haiper.ai 22
4.5 Závěr a implikace pro budoucí vývoj 23

5 Návrh řešení **24**
5.1 Funkce systému 24
5.2 Návrh uživatelského rozhraní 27

6 Implementace **30**
6.1 Základ webové aplikace 30
6.2 Generace scénáře 31
6.3 Implementace zvukových funkcí 32
6.4 Synchronizace délky videa se scénářem 33
6.5 Generace obrazů 33
6.6 Generace videa 35
6.7 Spojení videa do jednoho souboru 36
6.8 Spojení videa a zvuku 37

7 Testování a vyhodnocení **39**
7.1 Srozumitelnost a dostupnost systému 39
7.2 Rychlost tvorby reklam 39
7.3 Přizpůsobení různým typům reklam 40
7.4 Vyhodnocení 43

8 Závěr **45**

Literatura **46**

A Obsah přiloženého paměťového média **49**

B Plakát **50**

Seznam obrázků

3.1 Architektura modelu Transformer [32]. 9
3.2 Schématické znázornění architektury generativních adversariálních sítí (GAN), ilustrující spolupráci mezi generátorem a diskriminátorem. 10
3.3 Příklady lidských portrétů generovaných pomocí StyleGAN. Převzato z: `https://github.com/NVlabs/stylegan` 11
3.4 Architektura variačního autoenkodéru (VAE), zobrazující proces převodu vstupního obrazu do latentního prostoru a jeho následnou rekonstrukci. . . 12
3.5 Proces přímé difúze, ukazující postupné přidávání Gaussova šumu k datům. Převzato z: [8] 13
3.6 Proces obrácené difúze, demonstrující postupné odstraňování šumu pro obnovu dat. Převzato z: [8] 13
3.7 Ilustrace příkladu generování obrázků z textu pomocí modelu Stable Diffusion. Převzato z: [1] 14
3.8 Počáteční modely převodu textu na video byly omezené v rozlišení, kontextu a délce. Převzato z: [26]. 15
3.9 Sekvence generovaných videí modelem Phenaki podle popisů. Scény zahrnují astronauta chodícího po Marsu, tancujícího, vedoucího psa a sledujícího ohňostroj, vždy odpovídající daným popisům. Převzato z: [33]. 16
3.10 Ukázky v rozlišení 576×1024. V horní části: vzorky převodu obrazu na video (podmíněno nejlevějším snímkem). V dolní části: vzorky textu převedeného na video. 18
3.11 Ukázka různorodých výstupů modelu Lumiere: od generování videa z textu a obrázků, přes stylizovanou generaci až po video inpainting. 19

4.1 Snímek obrazovky Synthesia Studio. 21
4.2 Uživatelské rozhraní Pictory.ai zobrazující hlavní funkce pro transformaci textů na videa, včetně možností pro import skriptů, článků, úpravy videí a přidávání vizuálního obsahu. Převzato z: `https://www.elegantthemes.com/blog/business/pictory-ai-review` 22
4.3 Platforma Designs.ai využívající umělou inteligenci pro vytváření log, videí a reklamních materiálů. 22

5.1 Schéma procesu generování video obsahu na základě uživatelského vstupu. . 25
5.2 Původní prototyp uživatelského rozhraní s jedním vstupním polem pro všechny údaje. 28
5.3 Vstupní formulář webového rozhraní systému pro generování videoreklam. . 29
5.4 Vizuální prezentace procesu generování v uživatelském rozhraní. 29

6.1 Uživatelské rozhraní webové aplikace s formulářem pro zadávání parametrů reklamy. 31
6.2 Uživatelské rozhraní zobrazující proces tvorby reklamního videa. 38

7.1 Obrázky generované systémem zobrazující meditující ženu s brýlemi pro virtuální realitu. 41
7.2 Příklad obrázku generovaného mým systémem, který ukazuje problémy s kvalitou a zkreslení. 41
7.3 Obrázky znázorňující ekologické a inovační strategie značky, které byly generovány mým systémem. 42
7.4 Obrázky v různých stylech generované systémem pro IT kurzy pro dívky. . 44

B.1 Plakát prezentující tuto práci. 50

Kapitola 1

Úvod

Reklama se stala zásadní součástí úspěchu jakéhokoli produktu či služby. Díky pokroku v digitálních technologiích a širšímu přístupu k internetu se otevírají nové příležitosti v marketingu. Zejména videoreklamy jsou zásadní, protože umí rychle získat a udržet pozornost diváků. Tato videa nabízejí značkám schopnost sdělit složité myšlenky způsobem, který je vizuálně poutavý a snadno pochopitelný pro široké publikum. Vzhledem k tomu, že lidé tráví stále více času na internetu, nabízí videoreklamy jedinečnou šanci zaujmout potenciální zákazníky tam, kde se nejčastěji pohybují.

I když mají videoreklamy spoustu výhod, pro malé a střední podniky může být jejich tvorba drahá a časově náročná. Náklady na najímání kvalifikovaných odborníků mohou být vysoké, zejména pro ty, kteří se snaží udržet krok v konkurenčním prostředí.

V této situaci mohou systémy umělé inteligence nabídnout řešení. Existuje mnoho systémů umělé inteligence, které generují text, obrázky a videa a jsou využívány v různých oblastech od kreativních průmyslů po technické aplikace. Textové generátory mohou pomáhat s psaním textů pro webové stránky, zatímco generátory obrázků a videí vytvářet vizuální obsah bez potřeby lidského umělce. V oblasti videoreklamy systémy umělé inteligence revolučně mění tvorbu scénářů a videí, umožňují rychle generovat scénáře na základě zadaných témat, snižují náklady a zvyšují kvalitu reklamních videí.

V rámci mé práce byl vyvinut systém, který automaticky vytváří reklamní videa s pomocí neuronových modelů. Tyto modely generují scénáře a videa z dat zadaných uživatelem. Systém analyzuje informace, jako jsou cílové skupiny a speciální nabídky, aby vytvořil osobní a cílená reklamní videa. Použití systému zjednodušuje proces tvorby reklam a zajišťuje, že konečný obsah je přímo zaměřen a přitažlivý pro cílovou skupinu. Tento nástroj umožňuje značkám a marketérům provádět kampaně účinněji a šetřit zdroje, zejména čas a peníze.

V kapitole 2 jsou představeny základy reklamních videí, včetně klasifikace videí podle obsahu a délky. Dále jsou zkoumány tak důležité faktory, jako jsou cílové skupiny a unikátní nabídky, které jsou zásadní pro generování reklamního obsahu. Kapitola 3 je věnována popisu fungování různých neuronových modelů a procesům generování obrazů a videí, které budou následně využity pro vývoj systému. Kapitola 4 nabízí přehled existujících aplikací pro vytváření reklamních videí s využitím umělé inteligence. Rovněž jsou popsány jejich nedostatky. Kapitola 5 popisuje návrh systému, zatímco kapitola 6 se soustředí na jeho realizaci. Výsledky testování systému jsou prezentovány v kapitole 7.

Kapitola 2

Základy reklamního videa

Tato kapitola poskytuje podrobnou analýzu reklamního videa, která hraje důležitou roli při vývoji systému, který dokáže automaticky generovat cílené videoreklamy. Pochopení různých typů videoreklamy a faktorů, které ovlivňují její účinnost, jako je například délka videa, je nezbytné pro správnou úpravu nastavení systému tak, aby vyhovoval různým marketingovým a uživatelským požadavkům. Zvláštní pozornost je věnována určení hlavních komponent, jako je určení cílové skupiny a vytvoření unikátní nabídky, které jsou nezbytné pro správné nastavení vstupních dat systému a vytváření reklamních videí.

2.1 Typy reklamy

Jak je uvedeno v práci [41], reklamní video je druh reklamy, který využívá internetová videa jako komunikační kanál. Ve srovnání s jinými typy reklamy poskytuje významnou aktuálnost, široký dosah, rychlou zpětnou vazbu a široké možnosti dalšího rozvoje. Video může být užitečným nástrojem pro zvýšení rozpoznatelnosti značky a k přivedení návštěvníků na stránky sociálních sítí nebo webové stránky.

Podle obsahu

Video marketing se rozvíjí a obsahuje množství různých typů, z nichž každý plní specifické marketingové účely. Podle průzkumu publikovaného na portálu Wyzowl [37] se video recenze staly nejpopulárnějším typem, který používá 39 % marketérů. Za nimi následují vysvětlující videa (38 %), která objasňují složité koncepty, a prezentační videa (34 %), která poskytují podrobné informace o produktech nebo službách. Sociální mediální videa, také s podílem 34 %, jsou navržena tak, aby aktivně zapojila uživatele na platformách, kde tráví mnoho času.

K dalším důležitým typům videí patří demonstrační videa produktů (32 %), která ukazují, jak produkty fungují, reklamní videa (30 %), která jsou navržena k propagaci produktů nebo služeb, a prodejní videa (30 %) s cílem přesvědčit diváky k nákupu. Upoutávková videa (30 %) jsou navržena tak, aby vyvolala očekávání novinek, zatímco videa zákaznické podpory (28 %) poskytují pomoc po zakoupení produktu. Demonstrační videa aplikací účinně vizualizují data a funkce, a školicí videa pomáhají novým zaměstnancům nebo zákazníkům se rychleji zapracovat do firmy nebo seznámit se s produkty. Různé typy videí jsou vytvářeny s ohledem na různé fáze, kterými zákazníci procházejí.

Podle délky

Délka videa může významně ovlivnit jeho účinnost. Průzkum mezi marketéry ukázal, že nejúčinnější jsou krátká videa trvající mezi 30 a 60 sekundami, což potvrdilo 39 % dotazovaných. Následně nejlepší výsledky ukazují videa trvající 1-2 minuty (28 %), dále videa kratší než 30 sekund (18 %), videa mezi 2 a 3 minutami (10 %) a na posledním místě jsou videa delší než 3 minuty (5 %). Tyto statistiky ukazují, jak důležitá je správná délka videa pro zapojení publika [37].

2.2 Cílová skupina a jedinečná prodejní nabídka

Úspěch videoreklamní kampaně závisí na pochopení potřeb cílové skupiny, charakteristických vlastnostech nabízeného produktu nebo služby a na faktorech, které motivují diváka k podniknutí konkrétních kroků po zhlédnutí videa. Podrobná analýza těchto prvků významně přispívá k účinnosti v dosahování marketingových cílů [20].

Cílová skupina

Cílová skupina se definuje jako soubor osob, které mají vysokou pravděpodobnost stát se zákazníky produktů nebo služeb podniku [21]. Tato skupina je charakterizována společnými demografickými charakteristikami, jako jsou věk, pohlaví, geografická poloha, vzdělání a socioekonomický status [22]. Zaměření na rozpoznání cílové skupiny umožňuje lépe navrhovat marketingové strategie a rozpoznávat zásadní zákazníky. To vede k omezení zbytečných nákladů na marketing, který neoslovuje správné skupiny, a umožňuje tak zaměřit se na komunikaci s nejvhodnějšími zákazníky.

Podrobné určení věkových a geografických charakteristik cílové skupiny umožňuje nejlepší možné nastavení reklamních postupů. Přesné vymezení těchto údajů zvyšuje účinnost a příslušnost reklamních kampaní.

Jedinečná prodejní nabídka

V oblasti marketingu je koncept jedinečné prodejní nabídky základní strategií pro zdůraznění výhod značky nebo produktu oproti konkurenci.Jedinečná prodejní nabídka je charakterizována jako výrazný prvek produktu, který jej odlišuje od konkurenčních výrobků a poskytuje zákazníkům neopakovatelné výhody [6].

Výzkum, který provedl Talabi [30], se zaměřuje na to, jak co nejlépe využívat jedinečnou prodejní nabídku ke zvýšení účinnosti marketingových kampaní a zároveň ke zlepšení hodnoty pro zákazníky. Identifikuje důležité charakteristiky jedinečné prodejní nabídky. , mezi něž patří:

- **unikátnost** – jasně odlišuje produkt nebo službu od konkurenčních nabídek,
- **přesvědčivost** – motivuje zákazníky k nákupu,
- **jasně definovaná nabídka** – představuje zákazníkům přitažlivou alternativu.

Každá marketingová zpráva by měla zahrnovat specifickou nabídku pro zákazníky, která jasně ukazuje hodnotu unikátních vlastností produktu nebo služby. Tato nabídka by měla být jasně odlišná od toho, co nabízí konkurence dostatečně přitažlivá, aby přesvědčila potenciální zákazníky a rozšířila zákaznickou základnu.

Kapitola 3

Modely převodu textu na video

V dnešní době, kdy vizuální digitální obsah hraje zásadní roli v marketingu, vzdělávání a zábavě, je nezbytné automatizovat procesy vytváření a editace videí. Technologie umělé inteligence, jako jsou neuronové sítě a strojové učení, zrychlují proces videoprodukce a zlepšují kvalitu výsledného obsahu.

Významného pokroku bylo dosaženo ve vývoji modelů převodu textu na video. Proces tvorby video sekvencí na základě textových popisů je složitý. Na rozdíl od převodu textu na statický obraz je u videí potřeba, aby modely zvládly udržet plynulost a dynamiku obrazové sekvence.

Tato kapitola se zaměřuje na převod textových dat na obrázky a videa s využitím hlavních generativních modelů, jako jsou generativní adversariální sítě, variační autoenkodéry a difuzní modely. Jsou zde popsány základní principy každého modelu, jejich schopnost generovat realistické vizuální materiály z textů a kritéria pro jejich výběr podle požadavků na kvalitu, rychlost a rozmanitost výsledků. Porozumění těmto principům je důležité pro výběr nejvhodnějšího modelu, který bude základem pro vývoj systému.

3.1 Převod textu na obraz jako základ pro převod textu na video

Model text-to-image[1] je model strojového učení, který přijímá vstupní popis v přirozeném jazyce a vytváří obraz odpovídající tomuto popisu [3]. Tyto modely obvykle kombinují jazykový model, který převádí vstupní text do latentní reprezentace, s generativním obrazovým modelem, který na základě této reprezentace vytváří obrázek.

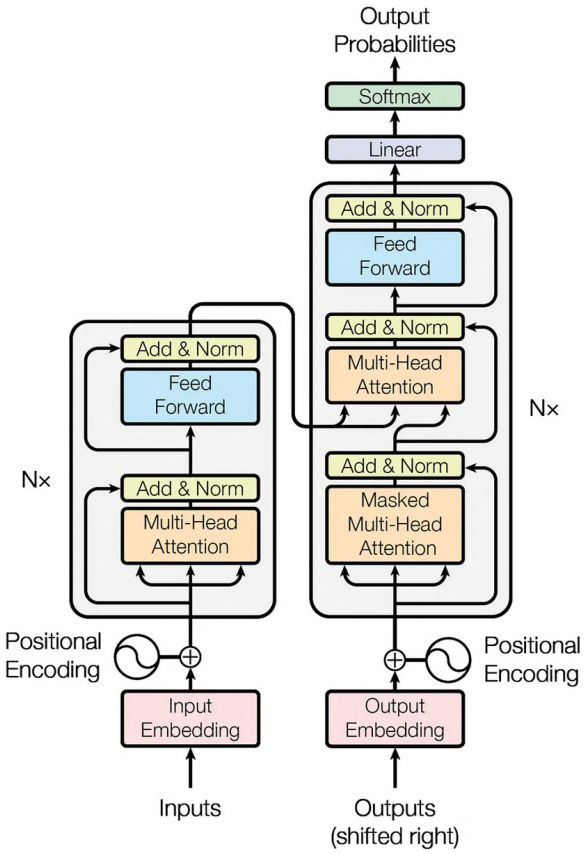
Jazykové modely

Jazykové modely hrají zásadní roli v oblasti zpracování přirozeného jazyka (NLP[2]). Díky jejich schopnosti hlubokého porozumění jazyku mohou syntetizovat a interpretovat text na úrovni srovnatelné s lidskou komunikací. Značný pokrok v této oblasti byl dosažen použitím metod hlubokého učení, které umožňují modelům učit se z velkých objemů neupraveného textu bez nutnosti manuální předúpravy dat. Jedním z hlavních průlomů v NLP se stala architektura Transformer, založená na mechanizmu pozornosti (attention mechanism) [32].

[1] Text-to-image – převod textu na obraz.

[2] Natural language processing (NLP) – zpracování přirozeného jazyka.

Tato architektura výrazně zlepšuje zpracování a pochopení dlouhých textových sekvencí, což napomáhá přesnějšímu modelování jazykových struktur.



Obrázek 3.1: Architektura modelu Transformer [32].

Příkladem modelu, který tyto pokroky využívá, je GPT-3[3]. Tento model rozšířil počet svých parametrů z 117 milionů na 175 miliard oproti své předchozí iteraci [16]. Díky tomu se podařilo výrazně zlepšit výkonnost modelů, což umožňuje zvládat složitější úlohy, jako je adaptace na nové situace a kontextové učení, což dává modelu schopnost uplatňovat získané znalosti v nových případech.

Všechny modely založené na architektuře Transformer, kterou lze vidět na Obrázku 3.1, mají několik komponent. Zásadní roli v této architektuře hrají transformační vrstvy, které mohou být dvou typů: kodér a dekodér. Původní architektura Transformer využívá oba tyto typy vrstev, zatímco některé pozdější modely se omezují pouze na jeden z nich.

Příkladem modelu používajícího pouze kodér je BERT [9], který se specializuje na porozumění textu a jeho analýzu. Na druhé straně, modely jako GPT, které používají pouze dekodér [42], jsou přizpůsobeny pro generování textu na základě daného kontextu.

V procesu převádění textu na obrazy jsou kodéry nezbytné pro počáteční kroky transformace textových popisů do vektorové formy, která může být dále zpracována pro vizuální reprezentace. Na začátku tohoto procesu kodéry analyzují vstupní text a převádějí jej do vektorových reprezentací, obsahující sémantické a kontextové informace z textu, které jsou důležité pro další zpracování v neuronových modelů zaměřených na generování obrazů.

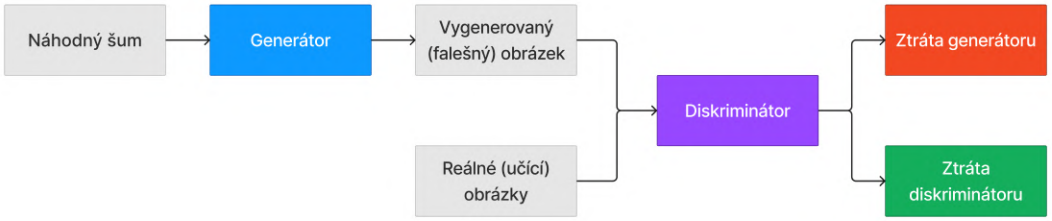
[3]Generative Pre-trained Transformer 3 (GPT-3) – generativní předtrénovaný transformátor 3.

Generativní umělá inteligence

Generativní umělá inteligence, která využívá principy generativních adversariálních sítí, variačních autoenkodérů a difúzních modelů, je zásadní součástí procesu převodu textu na obrázky a poté na video. Tyto modely, založené na učení z rozsáhlých datových sad, dokážou vytvořit nový, dosud neviděný obrazový obsah, který přesně a detailně odpovídá textovým popisům.

Generativní adversariální sítě

Generativní adversariální sítí (GAN[4]) jsou pokročilým algoritmem strojového učení, založeným na principu soupeření dvou neuronových sítí – generátoru a diskriminátoru [13], jejichž spolupráce je znázorněna na Obrázku 3.2.



Obrázek 3.2: Schématické znázornění architektury generativních adversariálních sítí (GAN), ilustrující spolupráci mezi generátorem a diskriminátorem.

Generátor G se učí generovat věrohodná data. Vygenerované případy se stávají negativními trénovacími příklady pro diskriminátor.

Diskriminátor D se učí rozlišovat nepravdivá data generovaná generátorem od skutečných dat. Diskriminátor penalizuje generátor za nevěrohodné výsledky.

Matematický model učení GAN lze popsat pomocí hodnotící funkce ve tvaru min-max hry [10]:

$$V(G,D) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))], \tag{3.1}$$

kde $\mathbb{E}_{x\sim p_{\text{data}}(x)}$ označuje očekávanou hodnotu přes rozdělení skutečných dat p_{data} a $\mathbb{E}_{z\sim p_z(z)}$ označuje očekávanou hodnotu přes vstupní šumové rozdělení p_z. $D(x)$ je pravděpodobnost, že x pochází z datového rozdělení podle diskriminátoru, a $G(z)$ je výstup generátoru z vstupního šumu z.

Cílem diskriminátoru je maximalizovat hodnotící funkci, což znamená, že se snaží přiřadit vysokou pravděpodobnost skutečným datům a nízkou pravděpodobnost datům generovaným generátorem. Naopak, cílem generátoru je minimalizovat logaritmický výraz $\log(1 - D(G(z)))$, což znamená, že se snaží, aby diskriminátor klasifikoval jeho generovaná data jako skutečná.

Tento matematický model umožňuje formulovat učení GAN jako hru s nulovým součtem, kde zlepšení generátoru vede k nutnosti zlepšení diskriminátoru a naopak, což vede k postupnému zdokonalování obou sítí.

[4]Generative adversarial networks (GAN) – generativní adversariální sítě

Generativní adversariální sítě představují nástroj pro celou řadu využití, od syntézy obrazů a simulací až po modelování v rozsáhlých datových prostředích. Jako jeden z významných příkladů GAN v oblastech umění lze uvést StyleGAN [17], model specializovaný na generování obrazů, který se vyznačuje vysokou kvalitou a širokým spektrem výstupů. Pro ilustraci ukazuje Obrázek 3.3 několik příkladů lidských portrétů generovaných pomocí StyleGAN. Model je známý svou schopností detailně simulovat různé styly a lze jej použít v různých kreativních a komerčních projektech.

Obrázek 3.3: Příklady lidských portrétů generovaných pomocí StyleGAN. Převzato z: https://github.com/NVlabs/stylegan

Variační autoenkodéry

Variační autoenkodéry (VAE[5]) jsou architekturou neuronové sítě, založenou na autoenkodérech, ale s integrovaným variabilním bayesovským přístupem [19]. Tato vlastnost odlišuje VAE od klasických autoenkodérů tím, že jim poskytuje možnost nejen rekonstruovat, ale i generovat nová data podobná těm, na kterých byly trénovány.

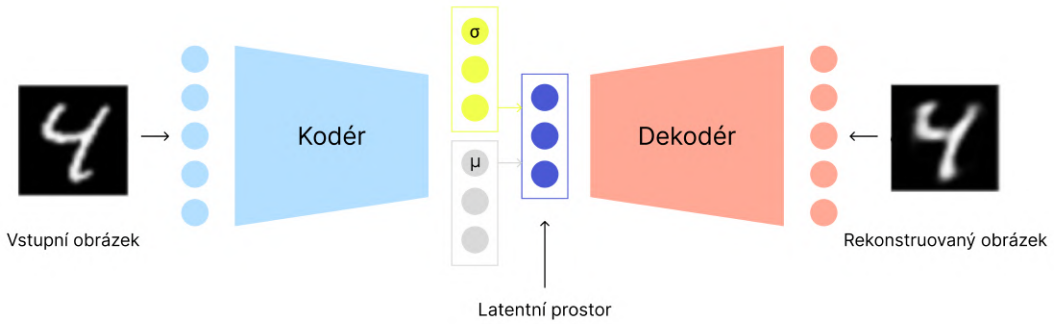
Kodér je zodpovědný za převod vstupních dat, jako jsou obrázky, do nízkorozměrného vektorového prostoru známého jako latentní prostor (viz Obrázek 3.4). Tento proces připomíná účinné stlačení informací, kde jsou základní vlastnosti dat extrahovány a uloženy ve zjednodušené formě.

Latentní prostor slouží jako kompaktní reprezentace informací obsažených ve vstupních datech. Umožňuje uchovávat vlastnosti dat ve výrazně zjednodušené podobě, ale stále dostatečně podrobné, aby bylo možné následně data rekonstruovat.

Dekodér pak přijímá tento stlačený vektor z latentního prostoru a převádí jej zpět do původního prostoru dat, čímž se snaží co nejpřesněji obnovit vstupní data. Úspěch této fáze je důležitý pro schopnost VAE nejen přesně rekonstruovat data, ale také vytvářet nová data s podobnými vlastnostmi, jaké měla původní trénovací sada.

Tato schopnost generovat nová data činí VAE cenným nástrojem pro řadu využití, od syntézy přes vytváření realistických obrazů až po pokročilé modelování v oblastech, jako jsou

[5]Variational autoencoder (VAE) – variační autoenkodér.



Obrázek 3.4: Architektura variačního autoenkodéru (VAE), zobrazující proces převodu vstupního obrazu do latentního prostoru a jeho následnou rekonstrukci.

hry, filmový průmysl nebo virtuální realita, kde může být generování nového, realistického obsahu užitečné.

Difúzní modely

Difúzní modely tvoří skupinu pokročilých generativních modelů, které si v posledních letech získaly značnou oblibu ve výzkumu strojového učení a umělé inteligence [8]. Základní myšlenka difúzních modelů spočívá v převádění dat do šumového stavu pomocí série kroků a následném obrácení tohoto procesu pro generování nových dat podobných původním.

Difúzní modely jsou založeny na třech zásadních procesech [7]:

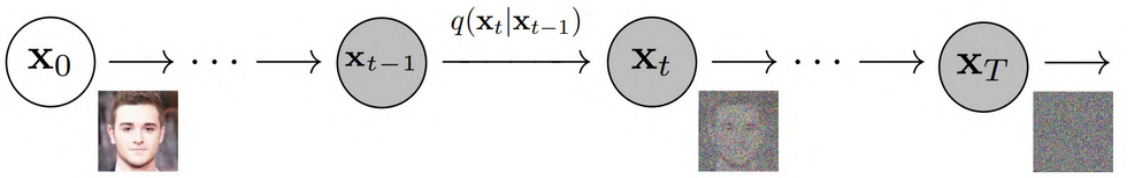
- **Přímá difúze:** Proces přímé difúze transformuje původní data x_0 do šumového stavu pomocí definovaného počtu diskrétních časových kroků. V každém kroku je do dat přidáno malé množství Gaussova šumu, postupně zvyšující neuspořádanost a snižující rozpoznatelnost původních vlastností dat. Tato postupná transformace je matematicky reprezentována rovnicí 3.2, která určuje přidávání šumu jako nezávislého a stejnoměrně rozděleného procesu v každém kroku t. Dále rovnice 3.3 určuje rozdělovací pravidla pro tento šum, kde $\mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$ představuje normální rozdělení s úpravou podle parametru β_t, udávajícího míru šumu v daném kroku. Tento proces je znázorněn na Obrázku 3.5, kde je zobrazeno postupné přidávání Gaussova šumu do původních dat.

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{3.2}$$

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \tag{3.3}$$

- **Trénování modelu:** Během trénovacího procesu se model učí obnovovat původní data ze zašuměných verzí. Toto se dosahuje pomocí variačního odvození, kde se model trénuje předpovídat, jaký šum byl přidán v každém kroku přímé difúze, a následně tento šum odstraňuje, aby obnovil data.

- **Obrácená difúze:** V generativním procesu model začíná s úplně náhodným šumem a postupně odstraňuje šum v procesu, který je podobný obrácení přímé difúze. Cílem

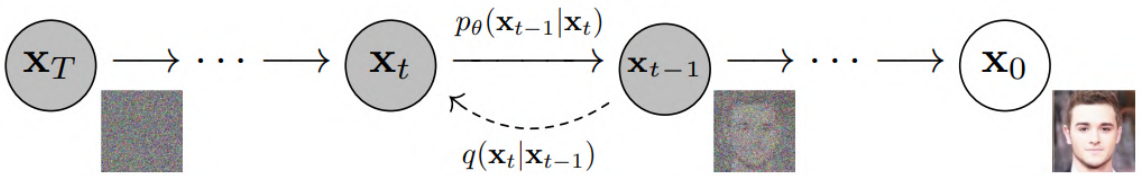


Obrázek 3.5: Proces přímé difúze, ukazující postupné přidávání Gaussova šumu k datům. Převzato z: [8]

je obnovit data, která jsou podobná tréninkovým datům. Tento proces je matematicky modelován pomocí rovnice 3.4, kde $p_\theta(x_{0:T})$ vyjadřuje rozložení dat, které model naučil, a kde θ jsou parametry modelu, které byly zdokonaleny během procesu učení. Rovnice 3.5 poté popisuje, jak model postupně 'odšumuje' data, používajíc k tomu podmíněné pravděpodobnostní rozložení s parametry závislými na θ, aby zrekonstruoval původní data z čistého šumu. Vizualizaci tohoto procesu lze vidět na Obrázku 3.6, který demonstruje postupné odstraňování šumu a obnovu dat.

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t \geq 1} p_\theta(x_{t-1}|x_t), \tag{3.4}$$

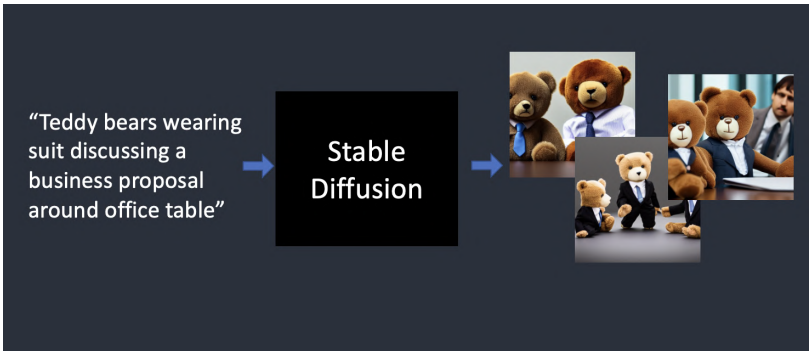
$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I), \tag{3.5}$$



Obrázek 3.6: Proces obrácené difúze, demonstrující postupné odstraňování šumu pro obnovu dat. Převzato z: [8]

Difúzní modely našly široké uplatnění v mnoha oblastech umělé inteligence [28] díky jejich schopnosti generovat vysoce kvalitní, realistická data:

- **Generování obrázků na základě textových popisů:** Jedním z nejdůležitějších použití difúzních modelů je generování obrázků [27], kde modely mohou vytvářet fotorealistické obrázky na základě textových popisů. Příklady zahrnují modely schopné generovat obrázky na základě popisů předmětů, scén nebo dokonce abstraktních pojmů. Jedním z příkladů je model Stable Diffusion [12], který umožňuje vytvářet detailní a působivé obrazy s vysokou úrovní koherence vizuálních detailů, jak je vidět na Obrázku 3.7.
- **Zlepšení kvality obrázků a videí:** Difúzní modely jsou také používány pro zlepšení rozlišení a kvality obrázků a videí tím, že obnovují detaily a zlepšují ostrost obsahu [8].



Obrázek 3.7: Ilustrace příkladu generování obrázků z textu pomocí modelu Stable Diffusion. Převzato z: [1]

Porovnání generativních modelů

V souvislosti s generováním obrazových dat každý z uvedených modelů – generativní adversariální sítě, difuzní modely, variační autoenkodéry – prezentuje specifické silné stránky a omezení. S ohledem na tři hlavní aspekty [38]: rychlé vzorkování, vysokou kvalitu vzorků a pokrytí režimů, rozdělení silných stránek a omezení modelů vypadá následovně:

- **Generativní adversariální sítě** jsou známé svou schopností generovat vzorky vysoké kvality, které jsou často nerozeznatelné od skutečných dat. Kromě toho mají tyto sítě také schopnost rychlého vzorkování. Nicméně mohou nastat problémy s pokrytím režimů, což vede k jevu zvanému kolaps režimu, kdy model generuje pouze omezenou rozmanitost vzorků.
- **Variační autoenkodéry** se vyznačují schopností rychlého vzorkování díky mapování do latentního prostoru a dobrou pokrytostí režimů, což umožňuje modelu vytvářet různé vzorky. Avšak občas mohou zaostávat v kvalitě generovaných vzorků, kdy vzorky nejsou tak ostré nebo detailní ve srovnání s jinými modely.
- **Difuzní modely** se vyznačují schopností generovat vzorky vysoké kvality s výjimečnými detaily a realističností. Mají také dobrou pokrytost režimů, což modelu umožňuje generovat velmi různé obrazy. Nicméně, proces generování vzorků může být pomalejší ve srovnání s generativními adversariálními sítěmi a variačními autoenkodéry, což omezuje jejich schopnost rychlého vzorkování.

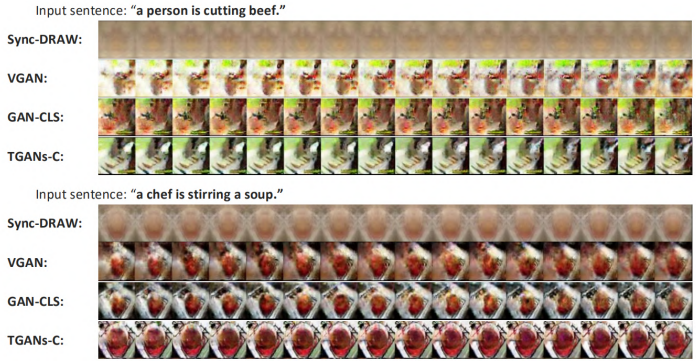
Pro systém určený k generování reklamních videí je zvláště důležitá kvalita a rozmanitost generovaných obrazů, což může zásadně ovlivnit výslednou uživatelskou spokojenost a úspěšnost reklamních kampaní.

3.2 Vývoj video technologií

Po prozkoumání modelů, které umožňují převádět text na obrázky – základní krok pro pochopení generativních technologií – nyní dochází k rozšíření jejich použití na oblast tvorby videa. Tato sekce se zaměřuje na vývoj metod od počátečních experimentů až po nejnovější pokročilé metody, které umožňují vytvářet videa z textových popisů s vysokým rozlišením a složitým obsahem. Pro tento účel byly využity informace z platformy Hugging Face [31], poskytující detailní analýzy a přehledy vývoje v této oblasti.

První vlna: Přístupy založené na GAN a VAE

Počáteční výzkum v oblasti tvorby videa z textu, který využíval techniky jako jsou generativní adversariální sítě (GAN) a variační autoenkodéry (VAE), byl přelomový. Příklady modelů z této doby, jako jsou Text2Filter [23] a TGANs-C [26], ukazují videa s omezeným rozlišením a dynamikou pohybu, jak je patrné z Obrázku 3.8. Tyto rané pokusy byly významně omezené v schopnosti generovat dlouhodobě kontextově soudržný obsah, což omezovalo jejich širší využití.



Obrázek 3.8: Počáteční modely převodu textu na video byly omezené v rozlišení, kontextu a délce. Převzato z: [26].

Druhá vlna: Přijetí architektur Transformer

Inspirována úspěchem velkých předtrénovaných modelů jako GPT-3 (pro text) a DALL-E[6] (pro obrázky), tato vlna začlenila architektury Transformer do generování videa. Modely jako Phenaki [33], Make-A-Video [29], NUWA [35], VideoGPT [39] a CogVideo [15] přijaly tento přístup, čímž zlepšily schopnost generovat delší a složitější video sekvence z textových popisů. Phenaki například umožňoval generování videí na základě posloupnosti popisů, vytvářející soudržnou dějovou linii. NUWA-Infinity [34] představila autoregresivní mechanismus schopný vytvářet nekonečný video obsah ve vysokém rozlišení. Pro ilustraci pokročilých schopností druhé vlny textově podmíněných modelů lze použít Obrázek 3.9, který demonstruje schopnost modelu Phenaki generovat videa podle specifických textových popisů. Výsledky jasně ukazují, jak mohou architektury Transformer sloužit k vytváření vizuálních sekvencí, což je zásadní krok v překonávání omezení raných GAN a VAE modelů.

Třetí vlna: Architektury založené na difúzi

Současná a nejpokročilejší vlna využívá modely založené na difúzi, které byly úspěšné v generování vysoce realistických obrázků. Tento přístup byl rozšířen do dalších oblastí, včetně videa. Modely jako Video Diffusion Models (VDM) [14] a MagicVideo [43] jsou předními příklady, přičemž MagicVideo zvyšuje účinnost tím, že pracuje v prostoru nižšího rozměru latentního prostoru. Model Tune-a-Video [36] ukazuje možnost vylepšení předtrénovaného modelu pro převod textu na obrazový obsah s využitím jediného páru textu a videa, což umožňuje upravit obsah videa při zachování původního pohybu. Následně se rozvíjející série modelů pro tvorbu videa z textových popisů zahrnuje systémy jako Video LDM [5],

[6] https://labs.openai.com/

Obrázek 3.9: Sekvence generovaných videí modelem Phenaki podle popisů. Scény zahrnují astronauta chodícího po Marsu, tancujícího, vedoucího psa a sledujícího ohňostroj, vždy odpovídající daným popisům. Převzato z: [33].

Text2Video-Zero [18], Runway Gen1 a Gen2 [11] a NUWA-XL [40]. Nedávno byly představeny novější modely, jako jsou Sora [25] a Stable Video Diffusion [4].

Model Stable Video Diffusion (SVD) je latentní video difúzní model určený pro generování videí ve vysokém rozlišení na nejmodernějším technologickém základě z textu do videa a z obrazu do videa.

Základním principem SVD je fungování v latentním prostoru, který odráží strukturu difuzních modelů obrazu a zároveň zahrnuje základní časové složky. Tyto složky zajišťují, že každý snímek videa je časově soudržný se svými sousedy, což má zásadní význam pro zachování jednotnosti v celé video sekvenci. Tento mechanismus umožňuje SVD zpracovávat složitou dynamiku videa, což z něj dělá vhodný nástroj pro tvorbu plynulých videí ze statických obrázků nebo textových popisů.

Trénink SVD je metodicky rozdělen do tří různých fází, z nichž každá je přizpůsobena postupnému zlepšování schopností modelu.

1. **Předtrénink na snímcích.** V této fázi se používá 2D difuzní model pro převod textu na obrázek k vytvoření počátečního porozumění různým objektům a scénám. Tento základní trénink je nezbytný k přípravě modelu na zvládnutí složitosti video obsahu.

2. **Předtrénink na videích.** Tato fáze zahrnuje trénink modelu na různorodých videodatech, což mu umožňuje přizpůsobit svůj latentní rámec tak, aby zvládal časové sekvence a specifické vlastnosti videa.

3. **Dokončení tréninku na videích s vyšším rozlišením.** Tato fáze se zaměřuje na zdokonalení výkonu modelu pomocí menší podmnožiny vysoce kvalitních videí s vyšším rozlišením. Je rozhodující pro zdokonalení modelu tak, aby vytvářel vizuálně působivá a dynamicky soudržná videa.

Specializované modely a techniky v rámci Stable Video Diffusion (SVD) ukazují, jak je tento systém přizpůsobivý a schopný, nabízejí řešení pro různé potřeby a díky důkladnému tréninku zlepšují kvalitu generovaných videí.

- **Model pro text do videa ve vysokém rozlišení** je doladěn na přibližně 1 milionu kvalitních videí s výrazným pohybem objektů a stabilním pohybem kamery. Doladění probíhá během 50 000 iterací v rozlišení 576×1024 s úpravou rozvrhu šumu pro dosažení nejlepší vizuální kvality. Model je navržen tak, aby z textových popisů vytvářel realistické video sekvence.

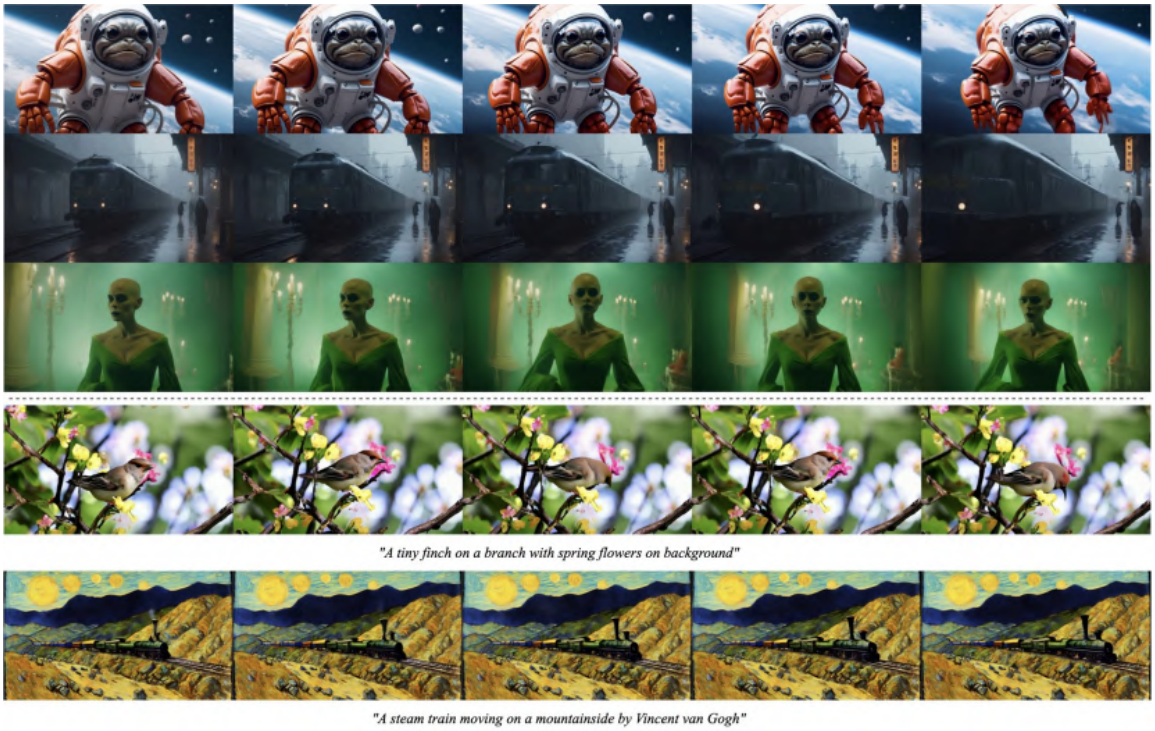
- **Model pro obrázek do videa ve vysokém rozlišení** převádí statické obrázky na video sekvence nahrazením textových vložek obrazovými vložkami z CLIP. Model se doladí ve dvou variantách: jedna generuje 14 snímků a druhá 25 snímků. Použití lineárně se zvyšujícího měřítka vedení mezi snímky zmírňuje artefakty a zajišťuje plynulost videa. Speciální tréninkové bloky (LoRAs) simulující pohyb kamery zlepšují přesnost napodobení různých pohybů, jako je posun, přiblížení nebo statické umístění, což zvyšuje věrohodnost a kvalitu video sekvencí.

Obrázek 3.10 přináší názorné příklady výstupů. Na horní části obrázku jsou prezentovány vzorky převodu obrazu na video, kde je video podmíněno krajním levým snímkem. Dolní část obrázku zase zobrazuje vzorky textu převedeného na video. Tato ukázka demonstruje, jak model SVD generuje plynulé video sekvence, které jsou v přímé souvislosti s podmíněným vstupem.

Podrobnější informace o modelu Stable Video Diffusion a o specifických technikách použitých v tomto systému lze nalézt v dokumentu [4].

Text2Video-Zero je model vytvořený výzkumnou skupinou Picsart AI Research ve spolupráci s významnými univerzitami, jako jsou Texaská univerzita v Austinu, Oregonská univerzita a Illinoiská univerzita v Urbana-Champaign [18]. Tento model je také difúzní a vyniká tím, že umožňuje vytvářet video obsah bez předchozího učení na video datech, což výrazně zjednodušuje a urychluje proces generování videa. Kromě vytváření videa z textu umožňuje model Text2Video-Zero také editaci videa. Například je možné vyměnit člověka ve videu za robota nebo změnit styl videa, aby vypadalo jinak. To činí model velmi praktickým pro různé video projekty.

Přesto model Text2Video-Zero má určitá technická omezení. Mezi ně patří potíže s udržováním časové konzistence a sekvence mezi snímky, což je důležité pro vytváření kvalitního



Obrázek 3.10: Ukázky v rozlišení 576 × 1024. V horní části: vzorky převodu obrazu na video (podmíněno nejlevějším snímkem). V dolní části: vzorky textu převedeného na video.

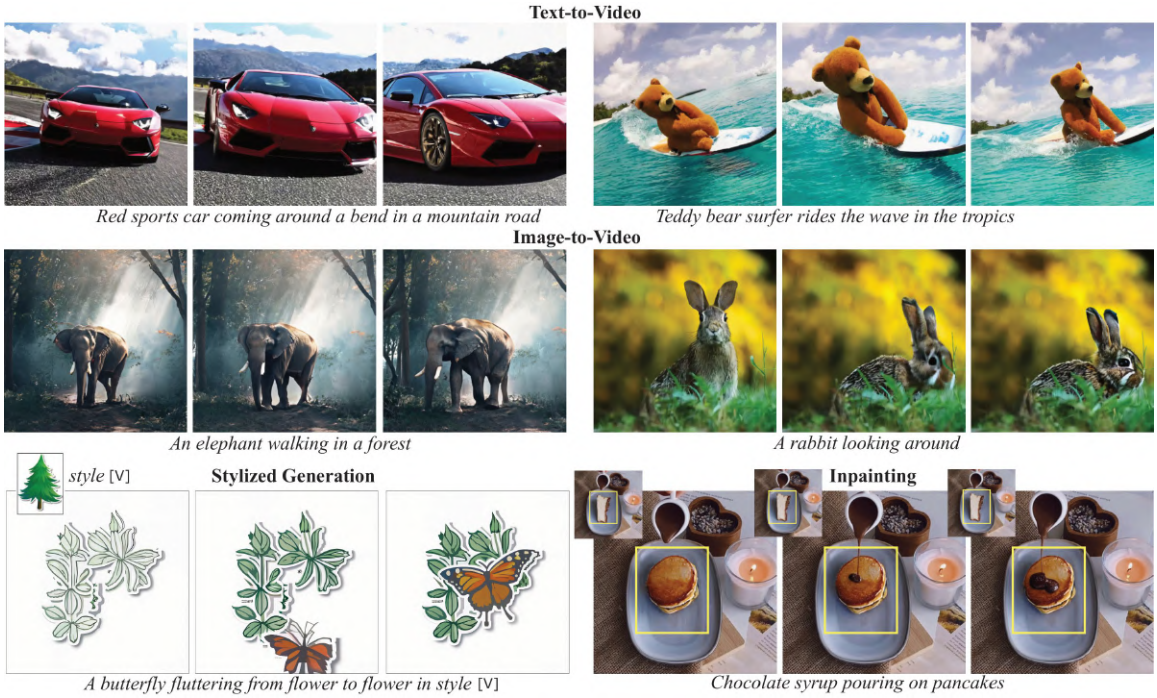
video obsahu. Model také čelí problémům při zpracování dynamických scén, což může ovlivnit realističnost a plynulost video sekvencí.

Lumiere nedávno představený Googlem, je nový difúzní model schopný generovat videa na základě textových nebo obrázkových popisů [2]. Tento model je založen na speciální architektuře nazvané Space-Time U-Net (STUNet). Díky této architektuře je model schopen vytvořit kompletní video v jediném kroku zpracování, což značně zvyšuje jeho účinnost. Navíc tato metoda zajišťuje, že v celém videu dochází k udržení časové souvislosti, což znamená, že všechny části videa jsou navzájem konzistentní a plynule na sebe navazují.

Lumiere umožňuje uživatelům nejen generovat videa, ale také je stylizovat a upravovat, což zahrnuje například změny ve vizuálním stylu nebo doplnění chybějících částí obrazu (video inpainting). Příklady toho lze nalézt na Obrázku 3.11. Díky tomu je model velmi flexibilní a nachází uplatnění v řadě různých aplikací, od osobních projektů až po profesionální mediální produkce.

Jako každá technologie, i Lumiere má určité omezení. Patří mezi ně například omezení na délku generovaného videa, což v současnosti činí maximálně pět sekund. Dalším problémem může být vysoká náročnost na výpočetní zdroje, což může omezit jeho použití na méně výkonných systémech.

Model Sora pro generování videí z textu, vyvinutý společností OpenAI a uvedený na trh 15. února, zásadně posunul hranice umělé inteligence v tvorbě videí [25]. Na rozdíl od svých současných konkurentů, jako jsou Runway Gen 2 a Pika, dokáže Sora tvořit 60sekundová videa z textových popisů, přičemž ukazuje pokročilou složitost a detailnost scén. Využívá



Obrázek 3.11: Ukázka různorodých výstupů modelu Lumiere: od generování videa z textu a obrázků, přes stylizovanou generaci až po video inpainting.

difuzní model, podobný systémům pro tvoření obrázků z textu, a zpracovává videa vykládáním prostorových a časových změn, přeměňující statický šum na srozumitelná videa. Avšak čelí výzvám v udržení fyzikální realističnosti, občas vytváří scény, které porušují zavedené fyzikální zákony, což je důsledek její závislosti na učení z dat místo jasného modelování fyziky.

Mezi omezení modelu patří nekonzistentní dodržování fyzikálních zákonů a vzájemných působení, což ukazuje, že jeho základem je rozpoznávání vzorů, nikoli hluboké porozumění fyzice. To často vede k nerealistickému zobrazení pohybových prvků a vzájemných působení ve vygenerovaných videích. Přesto vývoj Sory poukazuje na potenciál umělé inteligence ve tvorbě videí a naznačuje, že s rozvojem technologií lze očekávat přesnější simulace jevů reálného světa.

Je důležité zdůraznit, že na dobu psaní této práce nejsou modely Sora ani Lumiere veřejně dostupné. Oba projekty jsou stále ve vývoji a jejich přístupnost pro širší veřejnost zatím nebyla oznámena.

Kapitola 4

Přehled aktuálních nástrojů pro tvorbu videoreklamy

Pro vývoj systému pro tvorbu videoreklam s využitím umělé inteligence je důležité analyzovat a rozumět stávajícím technologickým řešením v této oblasti. V této kapitole je proveden přehled a analýza platforem umělé inteligence, jako jsou Synthesia.io[1], Pictory.ai[2], Haiper.ai[3] a Designs.ai[4], které již nabízejí nástroje pro automatizovanou tvorbu video obsahu. Prozkoumání funkčnosti těchto aplikací, jejich výhod a omezení umožní identifikovat zásadní aspekty, které je třeba zohlednit při vývoji nového systému.

4.1 Synthesia.io

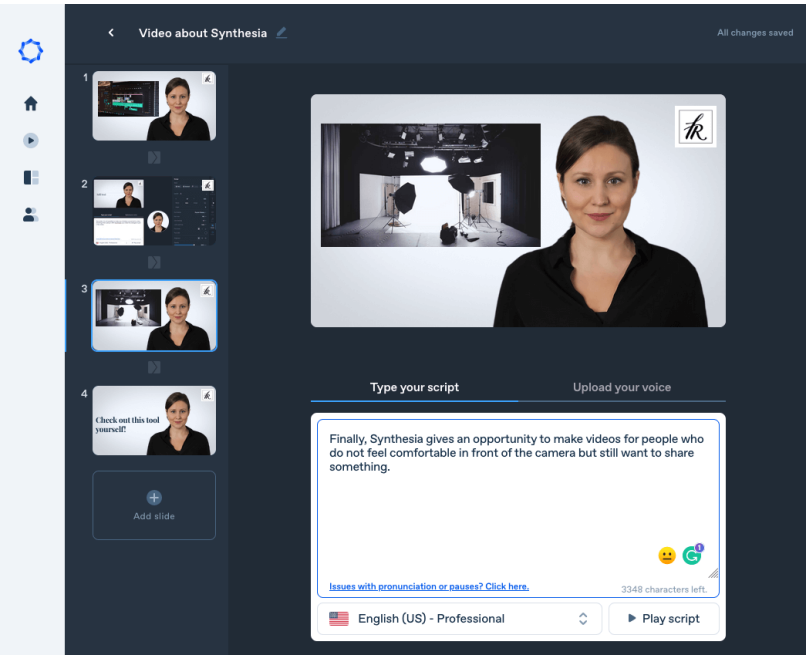
Synthesia.io představuje platformu pro tvorbu videí generovaných umělou inteligencí, která umožňuje vytváření obsahu s využitím virtuálních avatarů. Tato pokročilá technologie odstraňuje nutnost používat kamerové vybavení a herce, protože poskytuje nástroje pro rychlé a jednoduché vytváření videí. Uživatelé mohou zadat textový scénář, podle kterého Synthesia.io vytvoří video s odpovídajícím vyprávěním a pohyby avatara (viz obrázek 4.1). Platformu lze využívat v různých oblastech, od vzdělávacích materiálů po marketingové a komunikační strategie, což umožňuje uživatelům snížit výrobní náklady a současně zrychlit produkci obsahu.

Avšak, jak je běžné u mnoha technologických řešení, i Synthesia.io má svá omezení. Jedním z nich je potřeba, aby uživatelé sami tvořili text pro videa. Tento požadavek může být časově náročný a vyžadovat kreativní úsilí, což nemusí vyhovovat uživatelům, kteří hledají úplnou automatizaci procesu tvorby videa. Tato potřeba vlastní tvorby textu může omezit schopnost platformy plně uspokojit potřeby těch, kdo dávají přednost jednodušším řešením.

4.2 Pictory.ai

Pictory.ai je platforma umožňující snadnou tvorbu poutavých videí za pomoci AI bez nutnosti mít předchozí zkušenosti s úpravou videí. Pictory.ai umožňuje transformovat textové

[1] https://www.synthesia.io/
[2] https://pictory.ai/
[3] https://haiper.ai/
[4] https://designs.ai/



Obrázek 4.1: Snímek obrazovky Synthesia Studio.

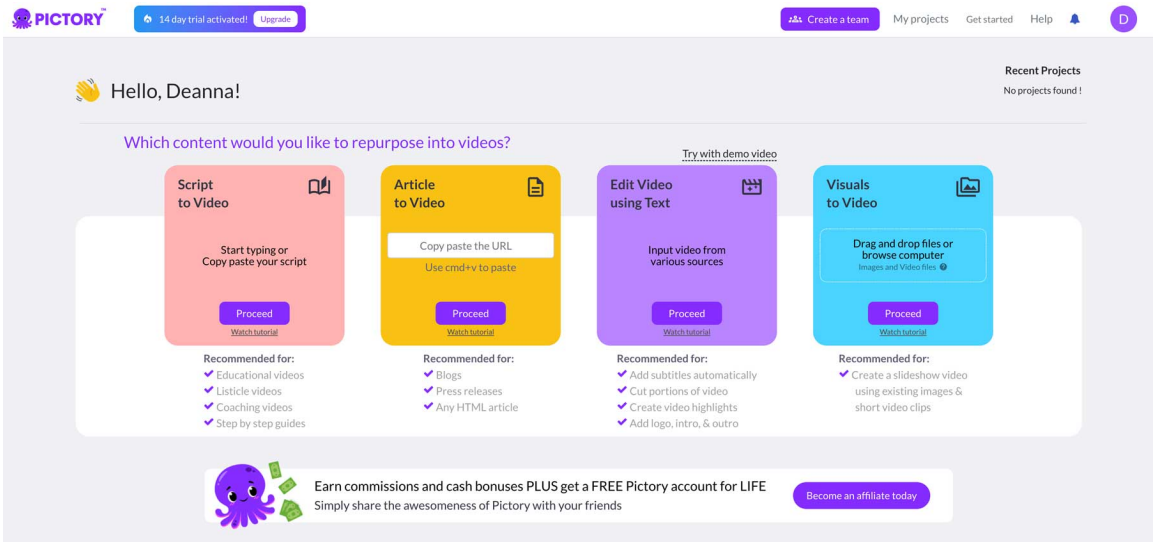
scénáře nebo blogové příspěvky na profesionálně navržená videa. Tento proces zahrnuje využití realistických hlasů na bázi umělé inteligence, vizuálních prvků a doprovodné hudby, což je dosaženo několika jednoduchými kliknutími. Na Obrázku 4.2 jsou znázorněny hlavní funkce a uživatelské rozhraní platformy, které poskytuje srozumitelný způsob, jak obsah převést do obrazové podoby.

I přesto, že platforma Pictory.ai nabízí mnoho výhod, má také několik omezení. Jedním z hlavních je, že kvalita vytvořených videí závisí na kvalitě vstupního textu. Pokud text není dobře strukturovaný nebo jasný, může to negativně ovlivnit finální video. Uživatelé musí také sami vytvářet a připravovat texty, které chtějí převést na video. Cena platformy může být dalším omezením, protože pro některé uživatele, zejména ty, kteří hledají pouze základní funkce, může být vnímána jako nevýhodná.

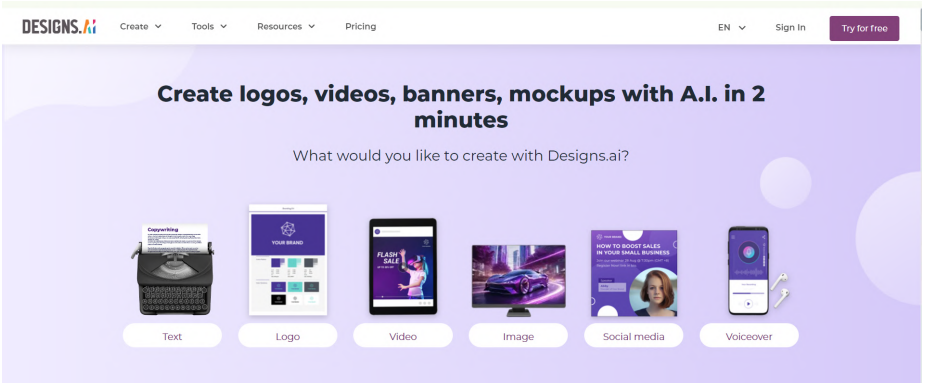
4.3 Designs.ai

Jak je vidět na Obrázku 4.3, platforma Designs.ai využívá umělou inteligenci k rychlému vytváření designů, jako jsou loga, videa a reklamní materiály. Stačí, když uživatelé zadají základní informace o svém projektu, a umělá inteligence zajistí zbytek, nabízí různé varianty designu na výběr.

Stejně jako jiné platformy, i Designs.ai pomáhá ušetřit čas a peníze ve srovnání s tradičními designovými procesy a činí profesionální designové služby dostupnější i pro menší firmy. Nicméně omezená kreativita a používání přednastavených šablon může vést k nedostatečné jedinečnosti výsledných designů, což může být nevýhodou pro značky, které usilují o výraznou identitu. Kromě toho, chybějící generátor obrázků založený na umělé inteligenci může omezit možnosti platformy v oblasti vizuálního obsahu.



Obrázek 4.2: Uživatelské rozhraní Pictory.ai zobrazující hlavní funkce pro transformaci textů na videa, včetně možností pro import skriptů, článků, úpravy videí a přidávání vizuálního obsahu. Převzato z: https://www.elegantthemes.com/blog/business/pictory-ai-review



Obrázek 4.3: Platforma Designs.ai využívající umělou inteligenci pro vytváření log, videí a reklamních materiálů.

4.4 Haiper.ai

Haiper.ai je další nástroj zaměřený na vytváření videí s pomocí generativní umělé inteligence. Umožňuje rychle vytvářet krátké klipy trvající pouze 2 nebo 4 sekundy pomocí textových pokynů a obrázků. Program také obsahuje funkci pro úpravu videa, která umožňuje měnit barvu nebo styl videa. Jelikož je aplikace poskytována zdarma, uživatelé se mohou setkat s významnými čekacími dobami na generování a nahrané klipy jsou doprovázeny logem Haiper AI. Hlavním problémem je, že kvůli krátkosti videa působí spíše jako animované obrázky bez dynamiky, například pokus o vytvoření videa s vzlétající raketou způsobí, že raketa zůstává na místě.

4.5 Závěr a implikace pro budoucí vývoj

Po analýze existujících platforem pro tvorbu video obsahu s využitím umělé inteligence, jako jsou Synthesia.io, Pictory.ai, Designs.ai a Haiper.ai, je zřejmé, že tyto systémy nabízejí významné výhody v automatizaci a zjednodušení procesů výroby mediálního obsahu. Tyto systémy snižují potřebu technických dovedností a zkracují čas i náklady na výrobu. Každá z těchto platforem však má svá omezení, jako například závislost kvality výrobku na vstupním textu, používání standardních šablon, které snižují unikátnost produktu, nebo potenciálně vysoké náklady, které nemusí být oprávněné pro uživatele hledající pouze základní funkce.

Proto při vývoji systému pro generování reklamních videí je třeba zvážit několik zásadních aspektů. Především je třeba se zaměřit na unikátnost vytvářeného obsahu. To umožní uživatelům vytvářet videa odpovídající jejich individuálním požadavkům. Dále je důležité implementovat funkci automatického generování scénářů, aby uživatelé nemuseli trávit čas jejich samostatným vytvářením. Kromě toho je důležité zajistit uživatelskou příjemnost a intuitivnost rozhraní. Systém by měl být přístupný uživatelům bez zkušeností s produkcí videa, což vyžaduje jasné a srozumitelné rozhraní a jednoduchost použití.

Kapitola 5

Návrh řešení

V kapitole 2 jsem se věnovala analýze hlavních typů videoreklam a zkoumala, jak délka reklamy ovlivňuje její efektivitu. V sekci 2.2 jsem se zaměřila na hlavní faktory, které mají vliv na úspěch reklamy. Zabývala jsem se také v kapitole 4 stávajícími aplikacemi pro tvorbu videoreklam, určila jejich slabé stránky a v kapitole 3 prozkoumala potenciál generativních modelů. Tyto shromážděné informace mají zásadní význam pro vytvoření nového systému. Na základě těchto dat jsem navrhla nový systém, který umožňuje automaticky generovat videoobsah na základě zadaných reklamních sdělení a základních scénářů.

5.1 Funkce systému

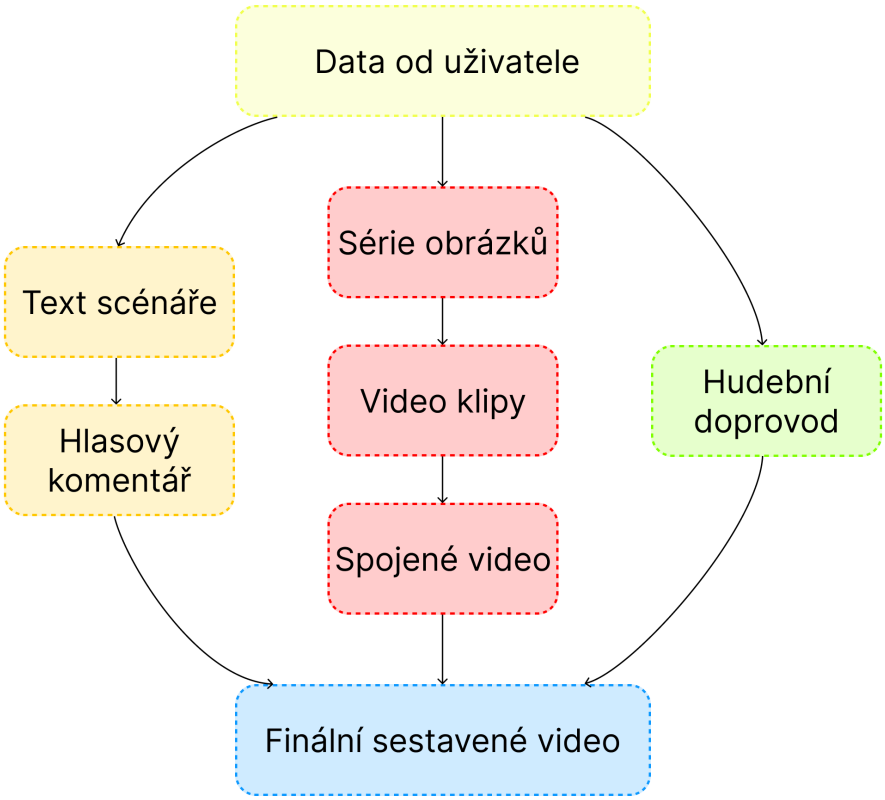
Systém se skládá z několika částí, které zahrnují: vstup informací potřebných pro generování reklamy uživatelem, generování scénáře, generování vizuálního obsahu, tvorbu zvuku a tvorbu finálního videa. Konkrétní kroky celého procesu generování jsou vizualizovány na Obrázku 5.1.

Generování scénáře

Jak bylo uvedeno po prozkoumání stávajících aplikací pro tvorbu videoreklam, mnohé z nich postrádají funkci automatického generování scénářů, což často nutí uživatele, aby si scénáře vymýšleli sami. Proto jsem se rozhodla zahrnout do mého systému funkci automatického generování scénářů. Tato funkce výrazně zjednodušuje proces tvorby videa.

Po analýze hlavních faktorů úspěšné reklamy v kapitole 2 jsem rozhodla, že uživatelé mého systému by měli mít možnost zadávat různé druhy údajů, aby systém mohl vytvořit scénáře, které přesně odpovídají marketingovým cílům a potřebám produktu. Jako vstupní data jsem zahrnula:

- podrobný popis produktu,
- definici cílové skupiny,
- jedinečnou prodejní nabídku,
- jasně formulovanou výzvu k akci (CTA),
- výběr nálady videa.



Obrázek 5.1: Schéma procesu generování video obsahu na základě uživatelského vstupu.

Považuji tento přístup k zadávání dat za velmi flexibilní, což umožňuje tvorbu široké škály videoreklam, které mohou cílit na různé skupiny diváků. Například, lze snadno přizpůsobit systém k vytvoření reklamy na oslavu určenou pro dospělé či děti, expedici do džungle pro dobrodruhy, nebo reklamu na pohádkové plyšové hračky pro malé děti. Tato flexibilita znamená, že systém může účinně sloužit různým marketingovým potřebám a cílům.

Pro automatické generování scénářů ve svém systému jsem se rozhodla využít model GPT-3.5 Turbo od OpenAI. Tento model je vhodný pro mou úlohu, protože dokáže vytvářet srozumitelný a přitažlivý text založený na podrobně specifikovaných vstupních datech. Navíc, GPT-3.5 Turbo je z finančního hlediska výhodnější ve srovnání s modelem GPT-4 Turbo, což je pro můj projekt nákladově přijatelné řešení.

Z analýzy prezentované v sekci 2.1, která zdůrazňuje účinnost reklamy trvající od 30 do 60 sekund, jsem dospěla k závěru, že nejlepší délka scénáře by měla být 3-4 věty. Tento rozsah je vhodný pro tvorbu videa odpovídající délky, poskytuje dostatek informací pro zaujetí a zapojení diváků, ale zároveň nesaturuje příliš mnoha informacemi.

Audio produkce

Pro zvýšení působivosti reklamy jsem se rozhodla, že každý scénář musí být namluven a doprovázen hudbou odpovídající náladě reklamního sdělení.

Generování hlasového komentáře

Jakmile je scénář hotový, systém přistupuje k dalšímu kroku: proměně textu ve zvukový záznam. Pro tuto úlohu jsem zvolila technologii převodu textu na řeč (TTS[1]) od OpenAI, protože tento model umožňuje transformovat napsané texty na přirozeně znějící řeč Tento krok umožňuje vytvořit plynulé a profesionálné podání, které bude sloužit jako doprovodný hlas v reklamním videu. Domnívám se, že tento proces nejenže činí video srozumitelnějším a přístupnějším pro širokou veřejnost, ale také mu dodává dynamiku.

Přidání hudebního doprovodu podle nálady

Po dokončení hlasového záznamu jsem se rozhodla, že systém by měl vybírat hudební doprovod odpovídající náladě, kterou uživatel zvolil na začátku. Možné nálady mohou být uklidňující, inspirující, napjaté, energické, nostalgické nebo romantické. Program automaticky vybírá vhodné skladby z předem připravené knihovny náhodně.

Generování vizuálního obsahu

Jednou z hlavních funkcí mého systému je tvorba vizuálního obsahu pro videoreklamy.

Generování obrázků

Ačkoli některé aplikace nabízejí výběr obrázků z předpřipravených knihoven, domnívám se, že pro dosažení větší originality a vizuální atraktivity je důležité využívat pokročilejší technologie. Namísto standardních obrázků jsem se rozhodla použít difuzní model Stable Diffusion od společnosti Stability AI, který je zmíněn v sekci 3.1. Tento model umožňuje vytvářet kvalitní a vizuálně atraktivní obrázky.

Transformace obrázků do videa

Pro převod generovaných obrázků na dynamické video sekvence jsem se rozhodla použít ve svém systému technologii Stable Video Diffusion, popsanou v sekci 3.2. Každý statický obrázek je převeden na čtyřsekundový videoklip s přidanou dynamikou a pohybem. Jsem přesvědčena, že tento způsob pomáhá udržet pozornost diváků.

Spojení všech video klipů do jednoho celku

Po transformaci jednotlivých obrázků do videí pomocí modelu Stable Video Diffusion, systém přistoupí k jejich spojení do jednoho uceleného videa.

Sestavení finálního videa

Tento krok považuji za důležitou součást svého systému, protože je nezbytný pro vytvoření soudržného a profesionálně vypadajícího produktu.

Kombinace všech prvků do finálního videa

Poslední fází procesu je spojení všech vytvořených prvků: video sekvence, hlasového doprovodu a hudby do jednoho celku. Tento proces vyžaduje přesnou synchronizaci zvukové

[1] Text-to-speech (TTS) – převod textu na řeč.

stopy s vizuálním obsahem a nastavení hlasitosti hudby tak, aby účinně doplňovala hlas, a ne jej přehlušovala.

Prezentace a stažení finálního videa

Jakmile je finální video připraveno, systém uživateli nabídne možnost jeho náhledu přímo v uživatelském rozhraní. To je důležité, aby se uživatel mohl ujistit, že video splňuje všechna jeho očekávání a požadavky. Pokud je uživatel s videem spokojen, může si jej stáhnout do svého zařízení pro další použití.

5.2 Návrh uživatelského rozhraní

Z hlediska použitelnosti mého systému jsem se rozhodla vytvořit webové rozhraní, které zajistí plynulou a intuitivní interakci s uživatelem. Během vývoje došlo k několika zásadním úpravám, které byly realizovány na základě zpětné vazby od uživatelů.

Původní prototyp a zpětná vazba uživatelů

Původní verze uživatelského rozhraní měla pouze jedno vstupní pole pro všechny informace, což způsobovalo, že uživatelé často nevěděli, jaké údaje mají vyplnit a jak to správně udělat. Po testování tohoto rozhraní s čtyřmi lidmi různých věkových skupin a profesí, včetně dvou univerzitních studentů, středoškoláka a podnikatele, jsem si uvědomila, že je nezbytné provedení změn.

Obrázek 5.2 ukazuje původní prototyp uživatelského rozhraní, který byl použit během těchto testů.

Hlavní problémy identifikované během testování:

- Nejasnost, co uživatelé mají do jednoho pole vkládat.
- Nedostatečná vizualizace procesu tvorby videa, což snižovalo srozumitelnost a použitelnost.

Upravené rozhraní

Na základě zpětné vazby, kterou jsem obdržela od uživatelů, jsem zásadně přepracovala původní prototyp a přetvořila jej do konečného prototypu, který lépe vyhovuje jejich potřebám:

Formulář pro zadávání údajů

Na hlavní stránce, kterou jsem navrhla, jsem nyní implementovala formulář. Struktura tohoto formuláře je podrobně zobrazena na Obrázku 5.3. Tento formulář je pečlivě strukturován do samostatných částí, kde je každá sekce přesně definovaná tak, aby poskytovala konkrétní informace – od podrobného popisu produktu, cílové skupiny, jedinečné nabídky až po formulaci výzvy k akci a definování emocionálního tónu připravovaného videa. Tento vstupní bod jsem navrhla s cílem zjednodušit proces předání potřebných informací.

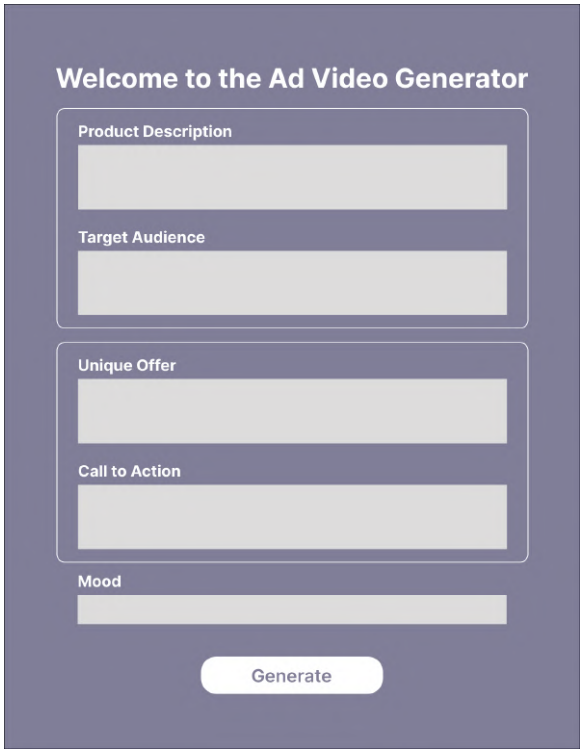
Vizuální prezentace procesu generování

Další stránka uživatelského rozhraní, kterou jsem přepracovala, nyní poskytuje úplné vizuální znázornění celého procesu generování. Považuji za velmi důležité, aby uživatelé mohli

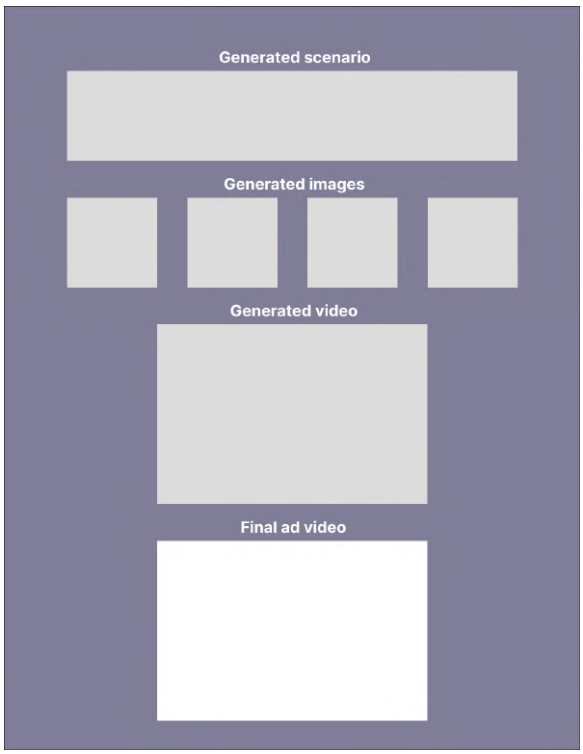
Obrázek 5.2: Původní prototyp uživatelského rozhraní s jedním vstupním polem pro všechny údaje.

sledovat každý krok procesu tvorby obsahu. Proto je nyní proces vytváření obsahu rozdělen do více kroků, jak je znázorněno na Obrázku 5.4. Každá fáze procesu je znázorněna interaktivním blokem. Tento přístup jsem zvolila po zjištění, že původní prototyp nebyl dostatečně srozumitelný a uživatelům chyběl přehled o stavu jejich projektů.

Při návrhu jsem také věnovala pozornost tomu, aby uživatelé co nejméně potřebovali technické znalosti. Interaktivní prvky jsem rozmístila tak, aby bylo zajištěno intuitivní ovládání a snadná identifikace, což pomáhá zlepšovat systém. Zvláštní pozornost jsem věnovala také logickému uspořádání obsahu, které uživatelům umožňuje přirozeně sledovat vývoj reklamy od počátečních fází až po finální vytvoření hotového videa.



Obrázek 5.3: Vstupní formulář webového rozhraní systému pro generování videoreklam.



Obrázek 5.4: Vizuální prezentace procesu generování v uživatelském rozhraní.

Kapitola 6

Implementace

V kapitole 5 byl představen návrh mého systému určeného pro vytváření reklamních videí. Tato kapitola poskytuje detailní pohled na technologie a procesy, které stojí za vývojem této aplikace, a popisuje jednotlivé funkce.

6.1 Základ webové aplikace

Pro vývoj webové aplikace, která umožňuje uživatelům generovat videoreklamy s využitím neuronových sítí, jsem se rozhodla použít kombinaci čtyř základních technologií: HTML, CSS, JavaScript a Python. Každá z těchto technologií hraje důležitou roli v procesu tvorby uživatelsky přívětivého a funkčního rozhraní, a také zajišťuje backendovou logiku pro zpracování dat a interakci s modely strojového učení.

HTML (HyperText Markup Language) je standardní značkovací jazyk, který slouží k vytváření struktur na webových stránkách. V projektu ho využívám k vytvoření základní struktury webové aplikace, včetně formulářů pro zadávání dat uživatelem, sekce pro zobrazení generovaného obsahu a tlačítek pro interakci s aplikací.

CSS (Cascading Style Sheets) definuje styl a vzhled webové stránky vytvořené pomocí HTML. V projektu používám CSS k nastavení stylů pro prvky rozhraní, jako jsou barvy, písma, což dělá uživatelské rozhraní atraktivní a srozumitelné.

JavaScript je používán k přidání interaktivity webovým stránkám. V kontextu mého projektu se stará o zpracování událostí, jako jsou kliknutí na tlačítka, odesílání formulářů a také asynchronní interakce se serverem pomocí AJAX. Tím umožňuje dynamickou aktualizaci obsahu stránky na základě dat získaných od serveru nebo vstupů od uživatele.

Python slouží jako serverový programovací jazyk, který řídí aplikační logiku. V projektu používám Python, který prostřednictvím frameworku Flask zpracovává HTTP požadavky, komunikuje s modely strojového učení a dalšími externími službami a následně odesílá zpracovaná data zpět klientovi. Vybrala jsem Python kvůli jeho výborné čitelnosti, bohaté knihovně nástrojů a snadné integraci s různými API, což je zásadní pro vývoj aplikace využívající strojové učení.

Tato kombinace technologií mi poskytuje vše potřebné pro kvalitní a rychlý vývoj webové aplikace.

Jak je ilustrováno na Obrázku 6.1, první část rozhraní aplikace představuje formulář pro zadávání údajů uživatelem. Tento interaktivní formulář, vytvořený s využitím HTML a CSS, umožňuje uživatelům definovat parametry videoreklamy, jako je popis produktu, cílovou skupinu, jedinečnou nabídku, výzvu k akci (CTA) a náladu reklamy. JavaScript je

použit k oživení interakce, zatímco backendová logika, včetně validace a zpracování dat, je řízena pomocí Pythonu s využitím frameworku Flask.

Obrázek 6.1: Uživatelské rozhraní webové aplikace s formulářem pro zadávání parametrů reklamy.

Tato struktura formuláře zajišťuje bezproblémový přenos informací od uživatele do systému, což je nezbytné pro automatické vytváření personalizovaných videoreklam.

6.2 Generace scénáře

Proces skriptování je důležitým prvkem mého systému a jednou z nejzajímavějších výzev, se kterými jsem se během vývoje setkala. Vše začíná přijetím dat od uživatele, která pokrývají vše od popisu produktu po náladu, kterou má reklama vytvořit. Tato data jsou odesílána na

server pomocí HTTP POST požadavku a v případě, že nejsou kompletní, server odpovídá chybovou zprávou.

Jak jsem zmínila v sekci 5.1, rozhodla jsem se pro generování scénáře využít model GPT-3.5 Turbo od OpenAI. Největší výzvou bylo přizpůsobit dotaz na API OpenAI tak, aby výsledný scénář byl koherentní a kontextově relevantní, zahrnoval informace zadané uživatelem a byl snadno srozumitelný. Pečlivě jsem navrhla strukturu textového dotazu tak, aby co nejlépe odrážel požadavky.

V kódu je tento dotaz formulován v angličtině, ale pro lepší porozumění je zde jeho překlad do češtiny:

```
Generujte scénář pro vypravěče, soustřeďte se pouze na slovní část scénáře
bez hudebního doprovodu, popisů scén nebo pokynů pro zvukové efekty.
Nálada tohoto scénáře by měla odpovídat {mood}.
Upravte scénář tak, aby obsahoval 3 věty.
Scénář by měl být založen na následujících informacích:
popis produktu: {product_description}, cílová skupina: {target_audience},
unikátní nabídka: {unique_offering}, výzva k akci: {cta}.
Prosím, poskytněte spojitý, plynulý narativ vhodný pro hlasový komentář
v reklamě.
```

Na základě tohoto dotazu model GPT-3.5 Turbo generuje text scénáře, který je serverem zpracován a odeslán zpět klientovi ve formátu JSON.

6.3 Implementace zvukových funkcí

Po úspěšné generaci scénáře následuje fáze generace zvukového doprovodu. Proto jsem vytvořila několik funkcí, které umožňují generování hlasových stopek, vybírání hudebního doprovodu podle nálady a jejich kombinaci do jednoho audio souboru.

Vytvoření hlasového komentáře

Jak jsem zmínila v sekci 5.1, k generování hlasového komentáře využívám technologii text-to-speech (TTS) od OpenAI, která transformuje text scénáře na přirozeně znějící řeč. Proces převodu textu na mluvené slovo zajišťuje funkce `generate_audio` a zahrnuje následující kroky:

1. Přijetí generovaného scénáře jako vstupního parametru.
2. Volání TTS (text-to-speech) API s tímto scénářem jako vstupem.
3. Uložení vygenerovaného audio souboru do specifikovaného adresáře na serveru.
4. Poskytnutí cesty k audio souboru klientovi pro další použití.

Výběr hudebního doprovodu

Druhá funkce, `generate_ad_music`, umožňuje uživatelům vybrat hudební doprovod, který odpovídá zvolené emocionální náladě videa. Hudba je řazena do složek podle nálady, a uživatel může specifikovat, jakou náladu hudební doprovod by měl odrážet. Systém poté vybere náhodnou skladbu z předem připravené složky.

Kombinace hlasového komentáře a hudebního doprovodu

Funkce `combine_audio`, kterou jsem implementovala, je určena pro finální úpravu zvukového složení. Tato funkce spojuje hlasový komentář a hudební doprovod do jediného audio souboru. Hlasitost hudby je nastavena na úroveň, která podporuje vyprávění, ale neovládá ho, obvykle na 20 % původní hlasitosti. Tento proces je řízen knihovnou `moviepy`, vhodnou pro zpracování s audio. Po načtení obou stop z uložených souborů na serveru jsou obě stopy sloučeny do jednoho výsledného audio souboru, který je pak převeden do formátu MP3 a uložen do definované složky na serveru pro další použití.

6.4 Synchronizace délky videa se scénářem

Jedním z úkolů, se kterými jsem se setkala při vývoji systému pro generaci reklamních videí, bylo dosažení synchronizace mezi délkou hlasového komentáře a vizuálním obsahem. Předchozí systém používal pevný počet obrázků a videí, což často neodpovídalo délce hlasového komentáře, což vedlo k nesouladu mezi zvukem a obrazem. To mohlo snižovat vnímání kvality reklamního sdělení.

K vyřešení tohoto problému jsem se rozhodla dynamicky stanovit počet obrazových prvků podle délky mluveného doprovodu. Tento přístup mi umožňuje přizpůsobit každé video přesně délce vyprávění, čímž dosahuji úplného souladu mezi zvukem a obrazem.

Délka celého videa se vypočítává podle vzorce:

$$\text{video\_duration} = \left\lceil \frac{\text{narrative\_audio\_clip.duration}}{3.5} \right\rceil \times 3.5 + 1 \tag{6.1}$$

Zde 3,5 sekundy představuje přibližný čas, který je potřeba k zobrazení jednoho videa generovaného z jednoho obrázku.

- Dělení délky audia na 3,5: Umožňuje určit minimální počet obrázků potřebných k pokrytí celé délky audia. Výsledné číslo je následně zaokrouhleno nahoru, aby bylo zajištěno, že počet obrázků bude dostatečný pro pokrytí celkové délky audia. Tento zaokrouhlený počet obrázků je využit v další fázi procesu, která zahrnuje generování obrazů. Detailnější popis této fáze bude podrobněji popsán v následující sekci 6.5.
- Zaokrouhlení nahoru: Zajišťuje, že video nebude ukončeno dříve než audio.
- Přidání jedné sekundy: Zajišťuje malý časový rezervní prostor pro případ, že by poslední obrázek nebo videoklip měl být zobrazen o něco déle před ukončením přehrávání.

Tímto způsobem je video přesně synchronizováno s délkou audia, což zlepšuje vnímání reklamy.

6.5 Generace obrazů

Jak jsem již zmínila v sekci 5.1, pro generování vizuálních prvků jsem zvolila pokročilý model Stable Diffusion od společnosti Stability AI. Bylo pro mě velmi důležité, aby model dokázal převést psaný text do přesných a detailních vizuálních obrazů. Vynikající kvalita výsledků a možnost přizpůsobit model konkrétním potřebám projektu byly rozhodujícími faktory při mé volbě.

V hodnocení nákladů na použití API Stable Diffusion jsem identifikovala značné ekonomické výhody. Vklad 250 korun za tisíc kreditů mi umožňuje generovat obrovské množství vizuálů, což přináší významnou úsporu ve srovnání s tradičními metodami tvorby obsahu. Nízká cena za realizaci jednoho videa, často nižší než půl koruny, zpřístupňuje rozšiřování projektů i s omezeným rozpočtem.

Navíc, minimální požadavky pro lokální využití modelu Stable Diffusion zahrnují počítač s 16 GB operační paměti a grafickou kartu NVIDIA s minimálně 8 GB VRAM. Abych systém zjednodušila, rozhodla jsem se používat API, což pomůže urychlit celý proces.

Příprava a odeslání dotazů

Proces přípravy a odesílání dotazů pro generování obrazového materiálu vychází ze stejného principu, který používám pro scénáře: je nutné definovat specifikace, které zahrnují detailní popis požadovaného obrazu. Tyto specifikace obsahují údaje jako kontext produktu, cílovou skupinu, vlastnosti produktu a atmosféru, kterou obrázek má vystihovat – stejně jako u generování scénáře. Tento popis se pak transformuje do dotazu, který je přizpůsoben pro využití schopností modelu Stable Diffusion, aby výsledný vizuální obsah byl nejen kvalitní, ale také co nejvíce relevantní.

Nejprve je nutné analyzovat uživatelský vstup specifikující požadovanou náladu nebo atmosféru, která má být na obrázcích zobrazena. Tento vstup se poté přiřazuje k předem definovanému slovníku `mood_descriptions` popisujícímu nálady, který zahrnuje klíčová slova spojená s určitými emocemi a atmosférami.

Na základě stanovené atmosféry systém vyhledá odpovídající popis nálad z tohoto slovníku, což umožňuje přizpůsobit dotaz tak, aby dosáhl nejlepších možných výsledků při generování obrazu. Tento postup zajišťuje, že vytvořené obrázky přesně odpovídají emocionálnímu zadání a vizuálním požadavkům uživatele.

Na základě specifikací a emocionálního kontextu, které jsem získala od uživatele, systém formuluje přesný dotaz pro generaci vizuálního obsahu. Tento dotaz, označený jako 'theme', je navržen tak, aby maximálně odpovídal požadavkům a využíval schopnosti modelu Stable Diffusion vytvářet obsah, který je nejen relevantní, ale i vizuálně přitažlivý. Struktura tohoto dotazu v reálném kódu je psána v angličtině, ale pro lepší pochopení je zde představena v češtině:

```
Ukažte {data['productDescription']}
v užití cílovou skupinou {data['targetAudience']} {mood_description}.
Zdůrazněte jeho {data['uniqueOffering']}.
Vizualizujte někoho, kdo se chystá {data['cta']}.
```

Takto sestavený dotaz zahrnuje všechny nezbytné informace pro účinné a cílené vizuální ztvárnění požadavků.

Interakce s API Stability AI a proces generování obrázků

Po definici a přípravě dotazu, označeného jako 'theme', se aktivuje funkce, která se nazývá `stability_generation`, jejímž cílem je transformace textového popisu na vizuální obsah. Tato funkce je zásadní pro překlad uživatelských zadání do konkrétních obrazových výstupů. Kód pro tuto funkci byl adaptován a modifikován na základě příkladu poskytnutého Jakubem Misilem na platformě LabLab, kde popisuje integraci modelu Stable Diffusion do existujících projektů [24].

Na začátku této funkce je inicializována instance `StabilityInference` s API klíčem. Toto nastavení zahrnuje aktivaci detailního režimu, který poskytuje podrobnější výstupy během generování obrázků, což je užitečné pro ladění a zlepšení procesu.

```
stability_api =
stability_client.StabilityInference(key=api_key, verbose=True)
```

Výpis 6.1: Inicializace instance `StabilityInference` s API klíčem.

Proměnná `stability_api` nyní obsahuje referenci na API klienta, který je připraven k odeslání dotazů pro vytváření obrázků.

Proces generování začíná nastavením systému, kde každý textový popis (prompt) je nezávisle zpracován, aby odpovídal uživatelským požadavkům. Obrázky se generují dle pečlivě navržených specifikací, které zahrnují následující kroky:

```
for i, prompt in enumerate(prompts):
    result = stability_api.generate(
        prompt=prompt,
        steps=40,
        sampler=generation.SAMPLER_K_DPMPP_2M,
        width=1024,
        height=576
    )
```

Výpis 6.2: Generování obrázků pomocí API klienta.

Každý textový popis je zpracován nezávisle, což zajišťuje, že model vygeneruje pro každý popis odpovídající obrázek.

Uložení a prezentace výsledků

Každý popis je zpracován individuálně, což zajišťuje, že výsledné obrázky přesně odpovídají požadavkům. Obrázky jsou ukládány lokálně a každý soubor je jednoznačně identifikován názvem obsahujícím časové razítko a pořadové číslo. Tyto uložené obrázky jsou poté připraveny k zobrazení v aplikaci a k dalšímu využití v procesu generování videa.

6.6 Generace videa

Další částí projektu je vytvoření videa, pro které jsem použila Stable Video Diffusion (SVD), který je popsán v této části kapitoly 3.2. Pro svůj systém jsem zvolila model SVD_XT, který je schopen generovat až 25 snímků na jedno video, což je významně více ve srovnání s původním modelem SVD, který vytváří pouze 14 snímků. Model SVD_XT je navržen tak, aby zvládl vyšší rozlišení 576x1024 pixelů na snímek, což umožňuje vytvářet videa s výrazně vyšší kvalitou obrazu. To je výhodné pro použití, která vyžadují vysokou úroveň detailů, jako jsou marketingové kampaně nebo vizuální prezentace produktů. Proces generování videa vychází z kódu, který byl převzat z GitHubu od Stability AI[1] a z projektu[2] a upraven.

[1] https://github.com/Stability-AI/generative-models
[2] https://github.com/sagiodev/stable-video-diffusion-img2vid

Požadavky a konfigurace modelu

Proces generování videa s využitím modelu Stable Video Diffusion (SVD_XT) vyžaduje významné množství výpočetního výkonu a paměti. Optimální výkon je možný na zařízeních s minimálně 16 GB operační paměti. Tento požadavek je zásadní pro zpracování vysokého rozlišení a datových struktur, které model během generování videa používá.

Proces generování videa začíná načtením a nastavením modelu prostřednictvím skriptu v jazyce Python, který využívá knihovnu PyTorch[3]. Nastavení modelu je definováno v souboru YAML, který obsahuje specifikace, jako je typ výpočetního zařízení (CPU/GPU), počet snímků a počet kroků vzorkování. Funkce `load_model` načítá model ze specifikovaného konfiguračního souboru `"svd_xt.yaml"`. Argument `"cuda"` určuje, že model bude spuštěn na GPU pomocí CUDA. Parametr `25` definuje počet snímků a parametr `30` počet kroků vzorkování pro dosažení vysoké kvality výstupu.

```
model = load_model("svd_xt.yaml", "cuda", 25, 30)
```

Výpis 6.3: Inicializace modelu s konkrétními parametry.

Zpracování obrázků a generace snímků

Proces generování snímků začíná po získání vstupního obrázku. Funkce `sample` generuje snímky pomocí předkonfigurovaného modelu. Přijímá cestu k obrázku (`input_path`), počet snímků (`num_frames`), počet kroků vzorkování (`num_steps`) a počet snímků za sekundu (`fps_id`), dále pohybovou transformaci (`motion_bucket_id`) a zařízení (`device`). Příklad volání funkce pro generaci snímků je následující:

```
output_paths = sample(
    input_path=input_path, resize_image=False, num_frames=25,
    num_steps=30, fps_id=6, motion_bucket_id=127, cond_aug=0.02,
    seed=23, decoding_t=4, device="cuda", output_folder="outputs")
```

Výpis 6.4: Ukázka generování snímků pomocí funkce `sample`.

Ukládání a export videí

Vygenerované snímky jsou následně skládány do finálního videa. Tento proces zahrnuje kompilaci jednotlivých snímků do video sekvence, která je uložena s unikátním identifikátorem, zahrnujícím časové razítko a pořadové číslo. Výsledné video je pak uloženo v definovaném výstupním adresáři `\outputs`, což usnadňuje jeho další používání.

6.7 Spojení videa do jednoho souboru

Dalším krokem bude spojení jednotlivých vygenerovaných videí do jednoho celku. K tomu použiji knihovnu `moviepy`, která poskytuje funkci `concatenate_videoclips`. Tato funkce umožňuje sekvenčně sloučit několik videí do jednoho dlouhého souboru.

```
def concatenate_videos(video_paths, output_path):
    clips = [VideoFileClip(video) for video in video_paths]
    final_clip = concatenate_videoclips(clips)
```

[3] https://pytorch.org/

```
final_clip.write_videofile(output_path, codec="libx264",
    audio_codec="aac")
```

Výpis 6.5: Funkce pro spojení několika videí do jednoho dlouhého souboru

Po spojení všech videí do jednoho souboru je finální video exportováno s použitím kodeku `libx264`, což zajišťuje dobrou kompatibilitu přes různé mediální platformy a zařízení. Finální video je uloženo v definované složce, což usnadňuje jeho archivaci a přístup pro další použití.

6.8 Spojení videa a zvuku

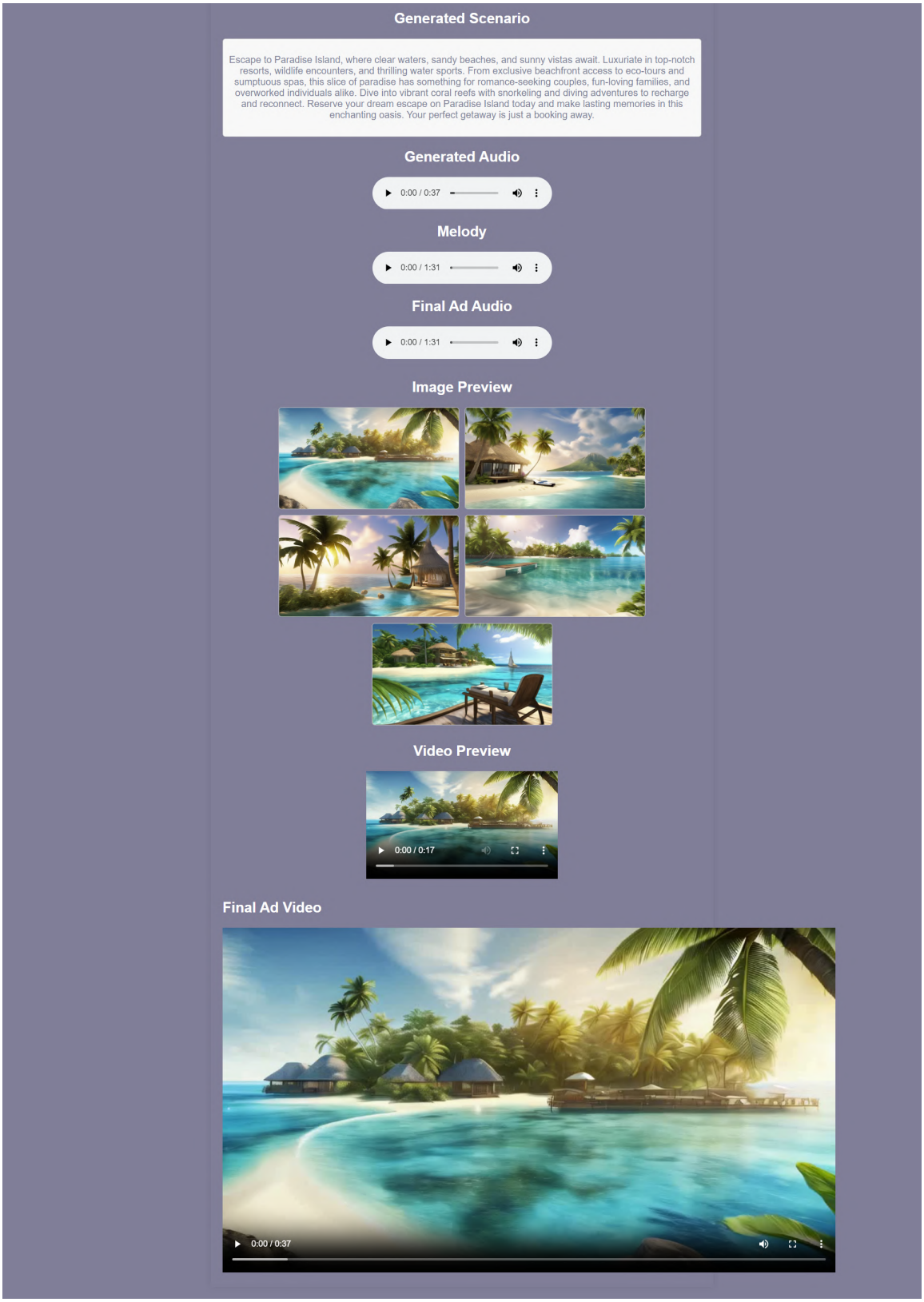
Pro spojení videa a zvuku využívám vlastní funkci `merge_video_and_audio` v Pythonu, která pracuje s knihovnou `moviepy`. Pomocí funkce `set_audio` z knihovny `moviepy` přiřazuji zvukovou stopu k videu.

```
final_clip = video_clip.set_audio(audio_clip)
```

Výpis 6.6: Přiřazení zvuku k videu pomocí funkce `set_audio`

Následně využívám funkci `write_videofile` k uložení finálního videa ve formátu MP4. Po úspěšném vytvoření je finální video uložené v složce `\outputs`, což umožňuje jeho snadné sdílení nebo vložení do webové stránky.

Celý tento proces, od generování scénáře po finální video, je vizualizován na uživatelském rozhraní, kde si může uživatel prohlédnout jak jednotlivé části, tak dokončené video. Příklad tohoto uživatelského rozhraní a jeho funkčnosti je možné vidět na Obrázku 6.2.



Obrázek 6.2: Uživatelské rozhraní zobrazující proces tvorby reklamního videa.

Kapitola 7

Testování a vyhodnocení

Tato kapitola je věnována důkladnému testování a hodnocení výkonu a účinnosti vytvořeného systému pro tvorbu reklamních videí. Zde analyzuji tři klíčové aspekty: srozumitelnost a dostupnost systému, rychlost tvorby reklamy a schopnost systému přizpůsobit se různým typům reklamy. V této kapitole se snažím identifikovat silné a slabé stránky systému a navrhnout možná zlepšení.

7.1 Srozumitelnost a dostupnost systému

Můj systém je navržen tak, aby byl jednoduchý a srozumitelný, což umožňuje uživatelům lehké pochopení a ovládání i bez hlubokých technických znalostí. Ačkoliv je rozhraní snadno srozumitelné, pro větší jistotu jsem přidala vysvětlující otázky u každého typu vstupu:

- Popis produktu: „Popište svůj produkt ...“
- Cílová skupina: „Kdo je vaše cílová skupina?“
- Unikátní nabídka: „Co dělá váš produkt jedinečným?“
- Výzva k akci: „Jakou akci by měl divák podniknout?“

Tento způsob zajišťuje, že každý krok procesu je jasně vysvětlen, což snižuje nebezpečí nedorozumění. Však je třeba zdůraznit, že můj systém je v současné době dostupný pouze v anglickém jazyce, což může některé uživatele omezovat. Považuji to za důležitý bod, který vyžaduje zlepšení v budoucích verzích, aby byl systém přístupnější širšímu okruhu uživatelů.

Pro ověření uživatelské přívětivosti systému jsem provedla průzkum mezi pěti osobami různého věku a profesního zaměření: dva studenti, školák, administrátor a podnikatel zvažující založení vlastní značky. Požádala jsem je, aby bez předchozích pokynů vyzkoušeli systém a zhodnotili jeho přehlednost a srozumitelnost. Účastníci byli vyzváni, aby samostatně zadali požadované údaje a postupovali dle návodů systému.

Výsledky ukázaly, že rozhraní bylo pro všechny uživatele snadno pochopitelné a dobře viditelné a jasně označené tlačítko pro pokračování v procesu účinně vedlo uživatele k správným činnostem.

7.2 Rychlost tvorby reklam

Při hodnocení použitelnosti mého systému jsem brala v úvahu také čas potřebný k vytvoření reklamního videa. Bez započtení času potřebného ke spuštění aplikace se průměrná doba,

kterou uživatelé potřebují k zadání údajů a kliknutí na tlačítko pro generování, pohybuje mezi 2 a 6 minutami. Tento časový rozsah závisí na množství informací, které uživatel plánuje zadat, a na čase potřebném k promýšlení odpovědí.

Po odeslání dat do systému začne proces generování částí, které zaberou nejvíce času. Samotné generování vizuálních prvků, kdy se každý obrázek převádí na video, může být časově náročné. Například generování pěti obrázků může trvat až 12 minut. Celková doba potřebná k vygenerování reklamy se tak může pohybovat od 15 minut do 40 minut i více, v závislosti na délce videa. Z této doby potřebuje uživatel k vložení informací o produktu v průměru pouze 4 minuty.

7.3 Přizpůsobení různým typům reklam

Při třetím testu jsem se rozhodla vyzkoušet schopnost systému generovat různé typy reklamních videí. Cílem bylo zjistit, jak dobře si systém poradí s různými požadavky specifickými pro jednotlivé typy reklam. Testovány byly následující typy: produktová reklama, firemní reklama, reklama na události.

Pro každý typ reklamního videa jsem připravila specifické vstupy, které napodobovaly reálné podmínky použití. Tento přístup mi pomohl otestovat, jak přesně dokáže systém generovat relevantní a poutavý vizuální a zvukový obsah pro každý z testovaných typů, a také identifikovat slabá místa v generování konkrétních typů reklam, které by mohly sloužit jako základ pro další vývoj a zdokonalení systém.

Ukázky vygenerovaných videí lze zhlédnout na disku ve složce `\videos`.

Produktová reklama

Účelem tohoto testování je ověřit schopnost programu generovat reklamní produktová videa. V tomto případě jsem vytvořila ukázku tvorby reklamy pro aplikaci virtuální reality ZenSpace VR, která je určena pro cvičení jógy a meditace.

Vstupy k vytvoření videa jsou následující:

- Popis produktu: ZenSpace VR je aplikace virtuální reality, která umožňuje uživatelům provozovat jógu a meditaci v poklidných virtuálních prostředích.
- Cílová skupina: Zaměřeno na mladé profesionály a technologicky zdatné jedince hledající úlevu od stresu prostřednictvím pohodlných domácích cvičení jógy a meditace.
- Unikátní nabídka: Kombinace špičkové technologie VR s personalizovanými, odborně vedenými sezeními jógy, které jsou přizpůsobeny individuálním zdravotním cílům.
- Výzva k akci: Připojte se k ZenSpace VR ještě dnes a zahajte cestu k duševnímu klidu doma!
- Nálada reklamy: Uklidňující / Klidná.

Systém vygeneroval následující reklamní scénář: „Opusťte chaos každodenního života a ponořte se do světa klidu s ZenSpace VR. Vytvořeno pro mladé profesionály a technicky zdatné jedince, kteří touží po úlevě od stresu, tato revoluční aplikace přenáší sílu jógy a meditace přímo k vašim prstům. Prožijte odborně vedené sezení v poklidných virtuálních nastaveních, které jsou přizpůsobené pro dosažení vašich zdravotních cílů. Připojte se k ZenSpace VR ještě dnes a zahajte cestu k duševnímu klidu doma!“

Obrázek 7.1: Obrázky generované systémem zobrazující meditující ženu s brýlemi pro virtuální realitu.

Program úspěšně vygeneroval reklamní scénář, který splňoval zadaná kritéria: cílová skupina, jedinečná nabídka a požadovaná nálada. Skript zdůrazňuje vlastnosti produktu a vyzývá k akci, což je základ úspěšné reklamy.

V průběhu testování systému bylo hodnoceno několik obrazových materiálů, které jsou relevantní pro téma reklamy virtuální reality určené pro meditaci. Obrázek 7.1 zobrazuje ženy meditující s brýlemi pro virtuální realitu, což jasně ilustruje zamýšlené použití produktu. Tyto obrázky jsou vhodné k ukázce funkcionalit aplikace a její atraktivnosti pro potenciální uživatele.

Přestože obrazové materiály jsou relevantní k tématu reklamy virtuální reality pro meditaci, v průběhu jejich analýzy jsem identifikovala několik nedostatků, které se týkají jednotvárnosti a kvality obrazů. Přílišná podobnost vizuálního stylu a kompozice u generovaných obrázků může vést k vizuálnímu přesycení, které má za následek snížení zájmu a pozornosti cílového publika.

Konkrétně, obrázek 7.2 ilustruje některé z těchto problémů s kvalitou generování, včetně nedostatku detailů a možného zkreslení, což může vést k zhoršenému vnímání produktu ze strany spotřebitelů.

Obrázek 7.2: Příklad obrázku generovaného mým systémem, který ukazuje problémy s kvalitou a zkreslení.

Firemní reklama

Dalším typem reklamy, kterou jsem se rozhodla otestovat, byla firemní reklama, jejímž cílem je vytvořit pozitivní vnímání značky a přilákat zákazníky, kteří sdílejí podobné hodnoty.

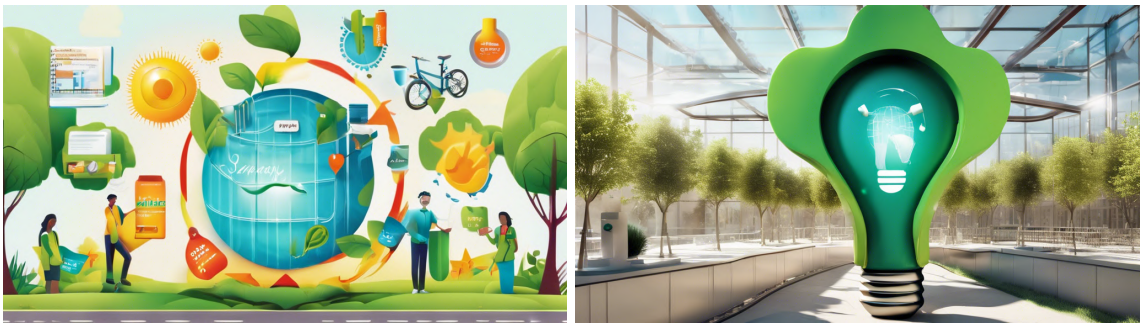
Jako příklad jsem vybrala reklamu na firmu, která klade důraz na inovace a udržitelnost ve všem, co dělá.

Vstupy k vytvoření videa jsou následující:

- Popis produktu: Usilujeme o inovace a udržitelnost ve všem, co děláme.
- Cílová skupina: Svědomití spotřebitelé hledající důvěryhodné značky se silnými firemními hodnotami.
- Unikátní nabídka: Závazek k šetrnosti k životnímu prostředí a kvalitnímu zákaznickému servisu.
- Výzva k akci: Připojte se k našemu hnutí za lepší budoucnost ještě dnes!
- Nálada reklamy: Radostné / Inspirující.

Systém vygeneroval následující reklamní scénář: „Ve světě, kde hodnoty ovlivňují naše rozhodování, jedna značka se vymezuje svým nasazením pro inovace a udržitelnost ve všem, co dělá. Pro svědomité spotřebitele, kteří vyhledávají důvěryhodné značky, nemusíte hledat dál než tento závazek k ochraně životního prostředí a vynikající péči o zákazníky. Připojte se k našemu hnutí za lepší budoucnost již dnes a buďte součástí komunity usilující o udržitelnější zítřek.“

Stejně jako při testování produktové videoreklamy, i při tomto testování systém vytvořil relevantní scénář se všemi hlavními prvky.



(a) Scéna zobrazující různé ekologické aktivity značky s použitím barevných symbolů jako jsou jízdní kola, recyklace a obnovitelné zdroje.

(b) Velká žárovka ve tvaru stromu umístěná v prosklené místnosti, obklopená skutečnými zelenými stromy, symbolizující inovace a ekologii.

Obrázek 7.3: Obrázky znázorňující ekologické a inovační strategie značky, které byly generovány mým systémem.

V tomto testování byly všechny obrázky generovány rozmanité a vhodné k tématu, v jasných barvách odpovídajících zvolené radostné náladě. Jako příklady lze uvést Obrázek 7.3a, který ukazuje scénu různých ekologických aktivit, a Obrázek 7.3b, na kterém je zobrazena velká žárovka ve tvaru stromu. Tyto obrázky nejen přitahují pozornost svou barevností a jasem, ale také účinně vyjadřují základní myšlenky hodnot značky: ekologie a inovace. Mezi nedostatky jsem zaznamenala občasnou nepřesnost ve vykreslení obrazů, což může snížit jejich vizuální účinnost.

Prezentační video

Další věcí, kterou jsem se rozhodla otestovat, je prezentace IT kurzů speciálně určených pro dívky. Vstupy k vytvoření videa jsou následující:

- Popis produktu: Interaktivní IT workshopy pro dívky, zaměřené na kódování, vývoj aplikací a kybernetickou bezpečnost.
- Cílová skupina: Dívky ve věku 12-18 let se zájmem o technologii.
- Unikátní nabídka: Mentorství od předních IT expertek a nejmodernějších technologických zařízení.
- Výzva k akci: Přihlaste se na další lekci a vydejte se na cestu za technologiemi! Počet míst je omezen. Přihlaste se ještě dnes a připojte se ke komunitě žen, které posilují pozici technologických lídrů.
- Nálada reklamy: Uklidňující / Klidná.

Systém vygeneroval následující reklamní scénář: „Přihlaste se na naše interaktivní IT workshopy určené pro dívky ve věku 12 až 18 let. Pod vedením předních odbornic na IT se můžete seznámit s kódováním, vývojem aplikací a kybernetickou bezpečností. Využijte této jedinečné příležitosti a odhalte svůj potenciál v nejmodernějších technologických prostorách. Počet míst je omezen – přihlaste se ještě dnes a připojte se ke komunitě posilující ženské technologické lídry!“

Stejně jako v předchozích případech, i v tomto testování byl scénář, který systém vygeneroval, relevantní a obsahoval všechny potřebné prvky, které byly důležité pro úspěšnou prezentaci kurzu.

Během testování jsem si všimla, že i když obrázky odpovídají tématu – dívky u počítačů, byly vytvořeny v různých stylech. To je vidět na Obrázku 7.4. Tento postup může být chápán dvěma způsoby: na jedné straně přidává kreativitu a rozmanitost vizuálnímu představení obsahu, což může být považováno za výhodu v marketingových a tvořivých projektech. Na druhé straně může používání více stylů v jedné kampani působit nevhodně, protože to může ovlivnit jednotnost a profesionalitu vizuální komunikace.

Kromě toho byly odhaleny vážné problémy s kvalitou tvorby lidských obrazů. Obličeje na některých obrázcích byly zkreslené, což vedlo k negativnímu dojmu z vizuálních materiálů. Tato zkreslení nejen snižují vizuální kvalitu, ale mohou také negativně ovlivnit vnímání značky nebo produktu spotřebiteli.

Podobné problémy byly zaznamenány i ve vygenerovaných videích. Špatná vizualizace, problémy s animací a zkreslení obličejů výrazně snižují celkovou kvalitu videomateriálů, což může odradit potenciální publikum a snížit účinnost videoreklamy.

7.4 Vyhodnocení

V této kapitole jsem se zaměřila na pečlivé hodnocení systému pro vytváření reklamních videí. Ačkoliv systém ukazuje vysokou srozumitelnost a dostupnost, což umožňuje i uživatelům bez hlubokých technických znalostí snadno a rychle vytvářet reklamní obsah, existují oblasti, které vyžadují zlepšení.

Podle mého názoru hlavním problémem je kvalita vizuálního obsahu, který systém generuje. Rozmanitost stylů a zkreslení lidských obličejů na obrázcích a ve videích může

Obrázek 7.4: Obrázky v různých stylech generované systémem pro IT kurzy pro dívky.

negativně ovlivnit vnímání vytvořených materiálů. To znamená, že v dalším vývoji je tedy nutné věnovat více pozornosti zlepšení textových dotazů pro generování obrázků, aby bylo dosaženo jednotné a vysoké kvality vizuální prezentace.

Navíc by bylo vhodné, aby systém podporoval více jazyků, což by umožnilo jeho širší využití a přístupnost pro rozmanitější skupinu uživatelů.

Kapitola 8

Závěr

Cílem práce bylo vytvořit systém, který pomocí neuronových modelů generuje reklamní videa na základě textových popisů. Tento systém má zjednodušit a urychlit tvorbu reklamních videí, snížit náklady a být vhodný i pro uživatele bez zkušeností v oblasti videoprodukce.

Při vytváření systému jsem se věnovala základům videoreklamy, prozkoumala různé typy a délku videí a detailně analyzovala faktory, jako jsou jedinečné nabídky a cílové skupiny, ovlivňující úspěšnost reklam. Výsledkem této analýzy bylo hlubší pochopení vhodných vstupních dat pro můj systém. Seznámila jsem se s principy fungování generativních modelů, abych při přípravě na automatizovaný systém pro produkci videoreklam mohla vybrat nejvhodnější nástroje pro převod textových dat na vizuální obsah. Volba modelu Stable Diffusion pro obrázky a Stable Video Diffusion pro videa byla motivována jejich schopností vytvářet vysokou kvalitu a fotorealistické výstupy. Z analýzy stávajících platforem pro tvorbu videoreklamy vyplynulo, že pro nový systém bylo důležité automatizovat tvorbu scénářů, zaručit vysokou kvalitu obsahu a poskytnout srozumitelné rozhraní.

Při tvorbě systému pro vytváření reklamních videí jsem kombinovala webové technologie a neuronové modely. Využila jsem HTML, CSS, JavaScript a Python s frameworkem Flask pro vytvoření rozhraní a zpracování dat. Uživatelé mohou zadat data, jako je popis produktu, jedinečné nabídky a cílovou skupinu. Systém automaticky vytvoří scénář pomocí modelu GPT-3.5 Turbo, vygeneruje video a připraví zvukové soubory s hudbou a generovaným hlasem. Výsledné video je pak dostupné k prohlížení a stahování přes webové rozhraní, což zajišťuje snadnou použitelnost pro uživatele.

Při testování systému pro generování reklamních videí byla důležitá kontrola funkčnosti, uživatelské přívětivosti a kvality obsahu. Testování zahrnovalo jak zpětnou vazbu od uživatelů, tak i analýzu generovaných reklamních videí na základě připravených vstupních dat. Pomocí uživatelů jsem testovala, jak pohodlné a intuitivní je uživatelské rozhraní webového aplikace. Také jsem zkoumala, jak dlouho trvá vytvoření reklamního videa a schopnost systému vytvářet různé typy reklam. Na základě výsledků testování plánuji zaměřit se na zlepšení kvality vizuálního obsahu a rozšíření podpory pro více jazyků, aby se posílila použitelnost systému pro širší mezinárodní publikum. Kromě toho mám v plánu přidat nové funkce, jako je možnost úpravy generovaného scénáře a umožnit uživatelům přidávat vlastní obrázky a hudbu ke svým reklamním videím. Tyto změny dle mého názoru výrazně zlepší uživatelskou zkušenost a umožní zvýšit jedinečnost při tvorbě reklamního obsahu.

Literatura

[1] Andrew. *How Stable Diffusion Works* [online]. 2024. Dostupné z: https://stable-diffusion-art.com/how-stable-diffusion-work/. [cit. 2024-04-12].

[2] Bar Tal, O.; Chefer, H.; Tov, O.; Herrmann, C.; Paiss, R. et al. *Lumiere: A Space-Time Diffusion Model for Video Generation.* 2024.

[3] Bie, F.; Yang, Y.; Zhou, Z.; Ghanem, A.; Zhang, M. et al. *RenAIssance: A Survey into AI Text-to-Image Generation in the Era of Large Model.* 2023.

[4] Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M. et al. *Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets.* 2023.

[5] Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W. et al. *Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models.* 2023.

[6] Blythe, J. *Essentials of Marketing.* 3rd. Financial Times Prentice Hall, 2005. 250 s.

[7] Chang, Z.; Koulieris, G. A. a Shum, H. P. H. *On the Design Fundamentals of Diffusion Models: A Survey.* 2023.

[8] Croitoru, F.-A.; Hondru, V.; Ionescu, R. T. a Shah, M. Diffusion Models in Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* Institute of Electrical and Electronics Engineers (IEEE), září 2023, sv. 45, č. 9, s. 10850–10869. ISSN 1939-3539. Dostupné z: http://dx.doi.org/10.1109/TPAMI.2023.3261988.

[9] Devlin, J.; Chang, M.-W.; Lee, K. a Toutanova, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* 2019.

[10] Ermon, S. *Generative Adversarial Networks* [online]. Dostupné z: https://deepgenerativemodels.github.io/assets/slides/cs236_lecture9.pdf. [cit. 2024-04-10]. Lecture 9, Stanford University.

[11] Esser, P.; Chiu, J.; Atighehchian, P.; Granskog, J. a Germanidis, A. *Structure and Content-Guided Video Synthesis with Diffusion Models.* 2023.

[12] Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J. et al. *Scaling Rectified Flow Transformers for High-Resolution Image Synthesis.* 2024.

[13] Goodfellow, I. J.; Pouget Abadie, J.; Mirza, M.; Xu, B.; Warde Farley, D. et al. *Generative Adversarial Networks.* 2014.

[14] Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M. et al. *Video Diffusion Models*. 2022.

[15] Hong, W.; Ding, M.; Zheng, W.; Liu, X. a Tang, J. *CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers*. 2022.

[16] Kalyan, K. S. A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, 2024, sv. 6, s. 100048. ISSN 2949-7191. Dostupné z: https://www.sciencedirect.com/science/article/pii/S2949719123000456.

[17] Karras, T.; Laine, S. a Aila, T. *A Style-Based Generator Architecture for Generative Adversarial Networks*. 2019.

[18] Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z. et al. *Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators*. 2023.

[19] Kingma, D. P. a Welling, M. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*. Now Publishers, 2019, sv. 12, č. 4, s. 307–392. ISSN 1935-8245. Dostupné z: http://dx.doi.org/10.1561/2200000056.

[20] Kluwer, W. *Effective Advertising Makes People Remember Your Name* [online]. 2024. Dostupné z: https://www.wolterskluwer.com/en/expert-insights/effective-advertising-makes-people-remember-your-name. [cit. 2024-04-10].

[21] Lake, L. *Consumer Behavior For Dummies*. Wiley, 2009. ISBN 9780470449837.

[22] Lake, L. *What Is a Target Audience* [online]. 2022. Dostupné z: https://www.thebalancemoney.com/what-is-a-target-audience-2295567. [cit. 2024-04-10].

[23] Li, Y.; Min, M. R.; Shen, D.; Carlson, D. a Carin, L. *Video Generation From Text*. 2017.

[24] Misilo, J. *How to integrate Stable Diffusion into your existing project* [online]. Dostupné z: https://lablab.ai/t/how-to-integrate-stable-diffusion-into-your-existing-project. [cit. 2024-04-21].

[25] OpenAI. *Video Generation Models as World Simulators* [online]. 2024. Dostupné z: https://openai.com/research/video-generation-models-as-world-simulators. [cit. 2024-04-12].

[26] Pan, Y.; Qiu, Z.; Yao, T.; Li, H. a Mei, T. *To Create What You Tell: Generating Videos from Captions*. 2018.

[27] Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C. a Chen, M. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. 2022.

[28] Saharia, C.; Chan, W.; Chang, H.; Lee, C. A.; Ho, J. et al. *Palette: Image-to-Image Diffusion Models*. 2022.

[29] Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J. et al. *Make-A-Video: Text-to-Video Generation without Text-Video Data*. 2022.

[30] Talabi, F. O. Making Slogans and Unique Selling Propositions (USP) Beneficial to Advertisers and Consumers. *New Media and Mass Communication*, 2012, sv. 3, s. 32. ISSN 2224-3267 (Paper), 2224-3275 (Online).

[31] Team, H. F. *Text to Video: The Next Frontier* [online]. 2024. Dostupné z: https://huggingface.co/blog/text-to-video. [cit. 2024-04-12].

[32] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L. et al. *Attention Is All You Need*. 2023.

[33] Villegas, R.; Babaeizadeh, M.; Kindermans, P.-J.; Moraldo, H.; Zhang, H. et al. *Phenaki: Variable Length Video Generation From Open Domain Textual Description*. 2022.

[34] Wu, C.; Liang, J.; Hu, X.; Gan, Z.; Wang, J. et al. *NUWA-Infinity: Autoregressive over Autoregressive Generation for Infinite Visual Synthesis*. 2022.

[35] Wu, C.; Liang, J.; Ji, L.; Yang, F.; Fang, Y. et al. *NÜWA: Visual Synthesis Pre-training for Neural visUal World creAtion*. 2021.

[36] Wu, J. Z.; Ge, Y.; Wang, X.; Lei, W.; Gu, Y. et al. *Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation*. 2023.

[37] Wyzowl. *Video Marketing Statistics 2024 (10 Years of Data)* [online]. 2023. Dostupné z: https://www.wyzowl.com/video-marketing-statistics. [cit. 2024-04-10].

[38] Xiao, Z.; Kreis, K. a Vahdat, A. *Tackling the Generative Learning Trilemma with Denoising Diffusion GANs*. 2022.

[39] Yan, W.; Zhang, Y.; Abbeel, P. a Srinivas, A. *VideoGPT: Video Generation using VQ-VAE and Transformers*. 2021.

[40] Yin, S.; Wu, C.; Yang, H.; Wang, J.; Wang, X. et al. *NUWA-XL: Diffusion over Diffusion for eXtremely Long Video Generation*. 2023.

[41] Zhang, H.; Mu, X.; Yan, H.; Ren, L. a Ma, J. A Survey of Online Video Advertising. *WIREs Data Mining and Knowledge Discovery (WIREs DMKD)*, 2023, sv. 13, č. 2, s. 3.

[42] Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X. et al. *A Survey of Large Language Models*. 2023.

[43] Zhou, D.; Wang, W.; Yan, H.; Lv, W.; Zhu, Y. et al. *MagicVideo: Efficient Video Generation With Latent Diffusion Models*. 2023.

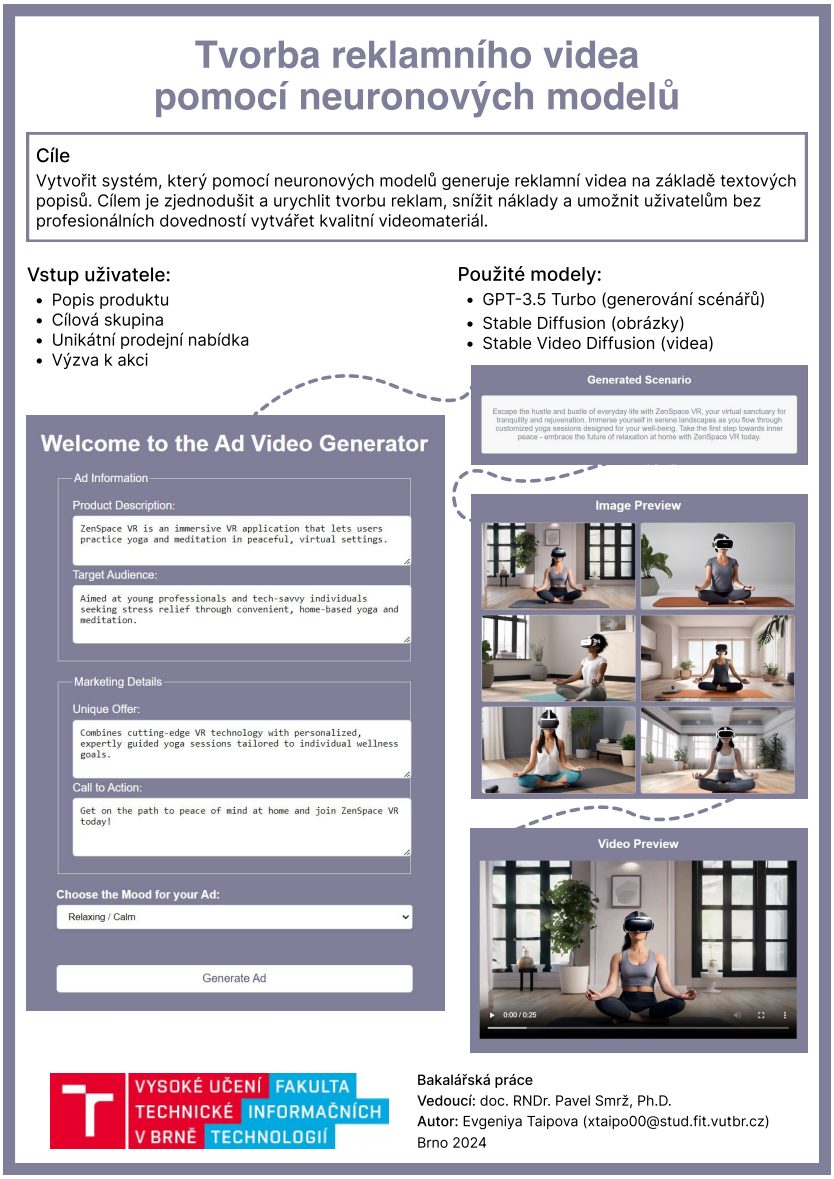
Příloha A

Obsah přiloženého paměťového média

- **my-app/** – zdrojové soubory aplikace a návod na spuštění v souboru `README.txt`.
- **doc/** – bakalářská práce v PDF.
- **doc/thesis/** – zdrojové soubory bakalářské práce.
- **videos/** – příklady vytvořených videí.

Příloha B

Plakát



Obrázek B.1: Plakát prezentující tuto práci.