Анализ трендов тем статей по генерации видео на данных arXiv.org за 2024 год

Подготовил: Кирюхов Григорий

4 июня 2025

Содержание

1 Задачи

Передо мной была поставлены следюущие задачи:

Выделить основные тренды в видео генерации в 2024 году.

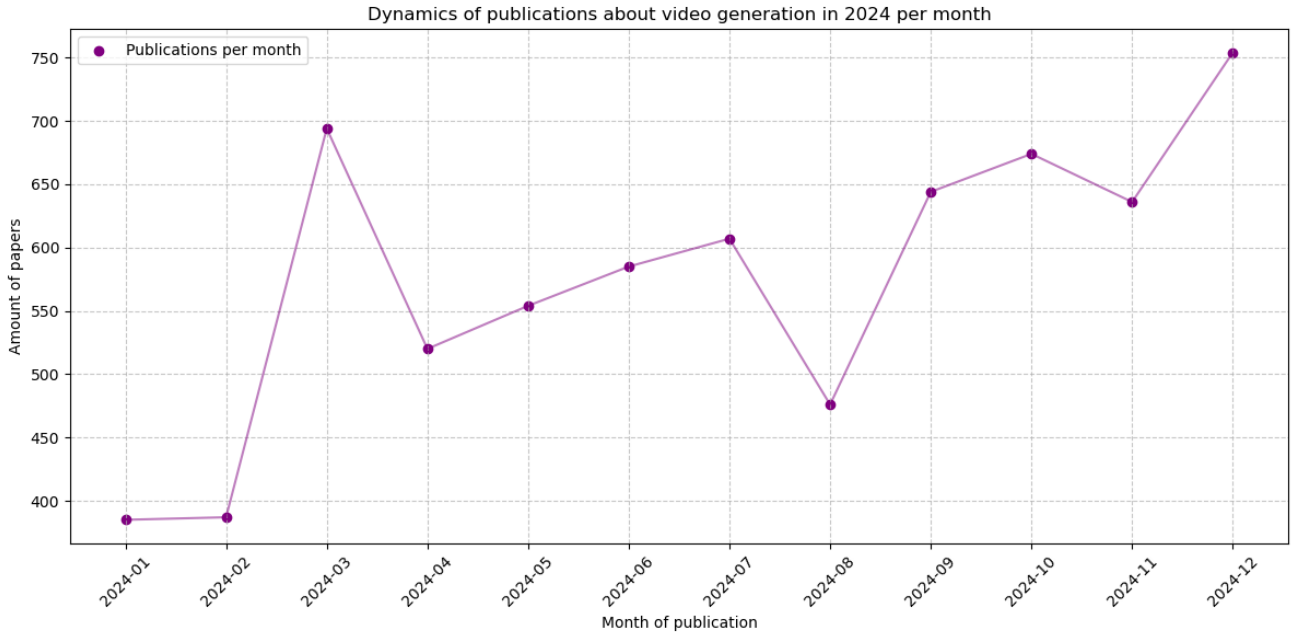
1. Получить метаданные статей с `arXiv` (любым способом) по теме “видео генерация”, выпущенных в 2024 году.
2. Написать код для выделения основных тем/трендов.
3. Будет плюсом, если каждая тема будет представлена связным словосочетанием/предложением
4. Представить визуализацию результатов.

2 Краткий обзор результатов

Была собрана база данных из 8870 статей по теме видеогенерации, из которых 6915 относятся к 2024 году, а остальные — к 2025. Выборка за 2025 год использовалась для дополнительной проверки статистической гипотезы о наличии тренда.

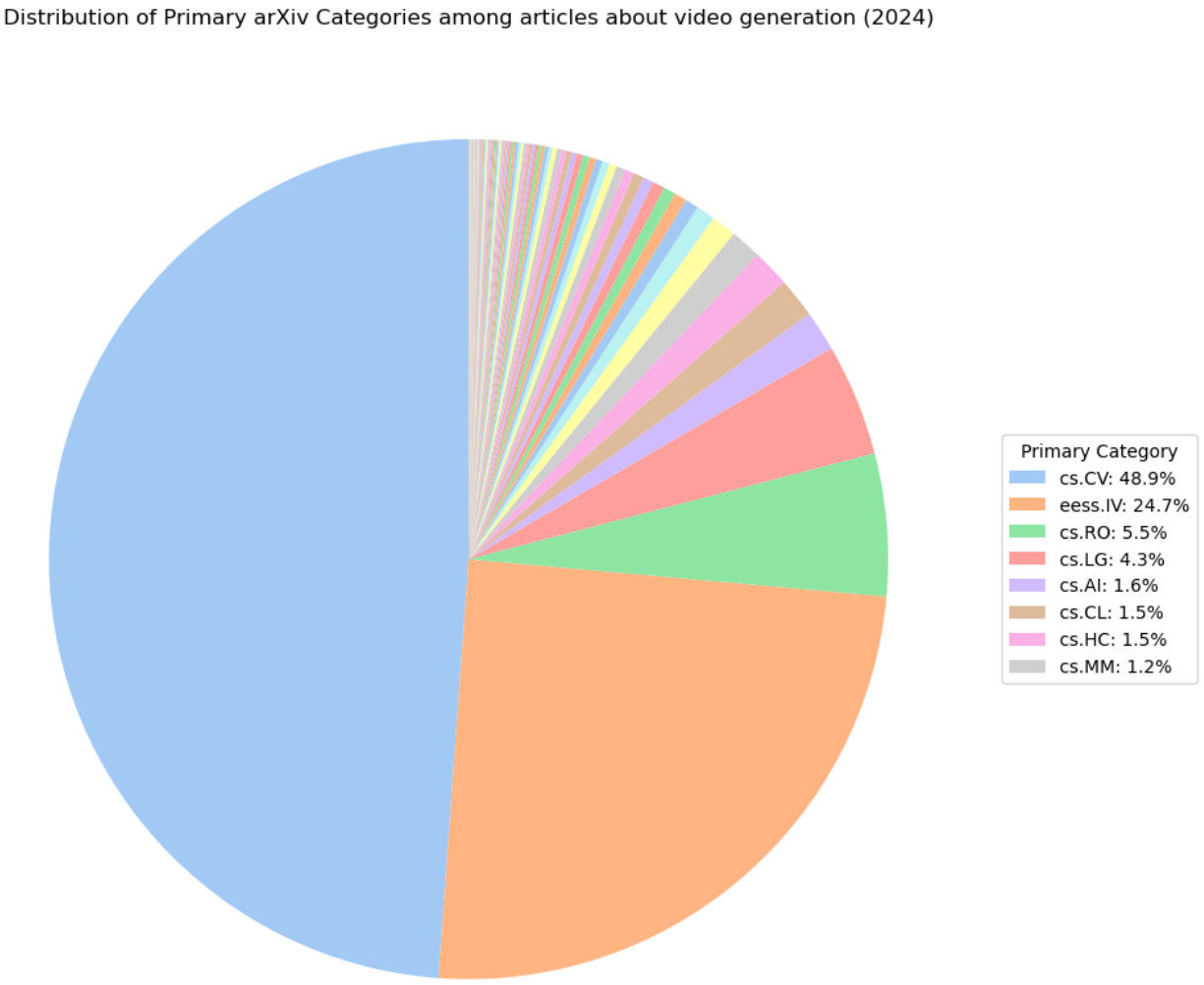
Оказалось, что ежемесячно выходит более 400 статей, и можно наблюдать некий положительный тренд в их появлении. Однако делать окончательные выводы пока рано, поскольку данные могут быть зашумлены.

Ниже приведён график динамики публикаций:



Также проведён анализ соответствия категорий. В `arXiv` у каждой статьи есть метка `primary_category`, определяющая её принадлежность к определённой категории. Более подробную информацию о метках можно найти на официальном сайте arXiv.org.

Ниже приведён barchart, иллюстрирующий распределение статей по основным категориям:



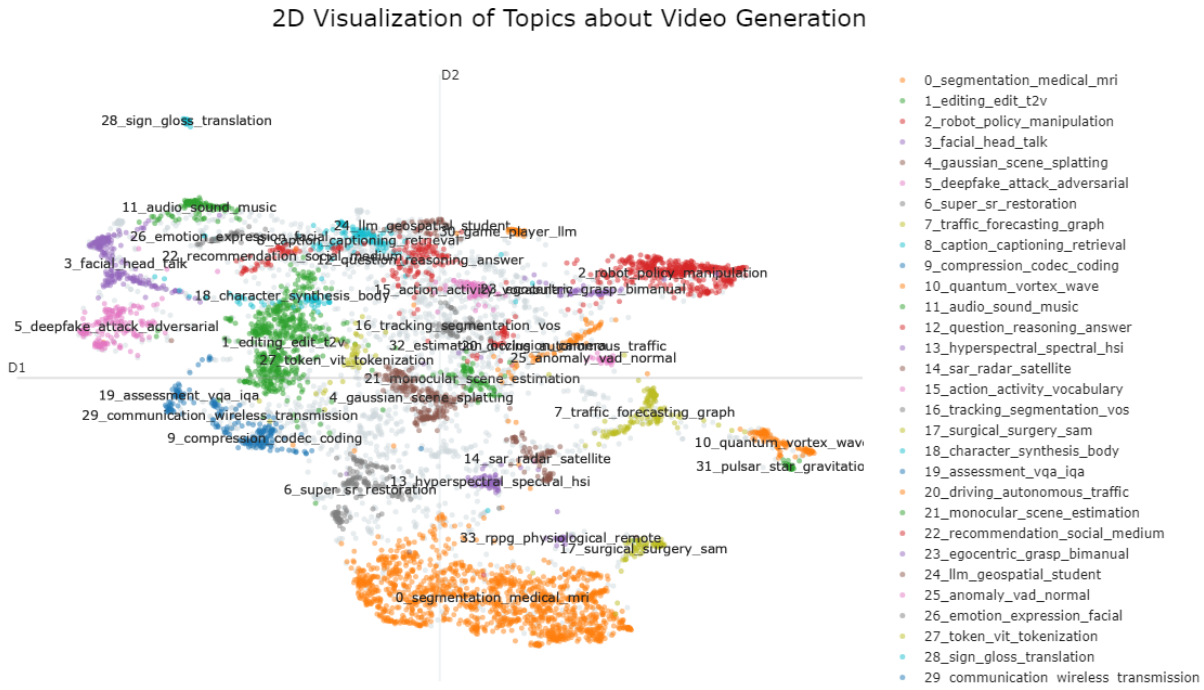
Как видно из диаграммы, топ-2 категории — `cs.CV` (Computer Vision and Pattern Recognition) и `eess.IV` (Image and Video Processing). Это свидетельствует об адекватности подхода к сбору данных. Следует отметить наличие большого количества аутлайеров в виде мелких категорий, что было учтено при выборе модели.

Для решения поставленной задачи была выбрана модель BERTopic (обоснование выбора приведено в соответствующем разделе). После подбора гиперпараметров были получены следующие метрики качества кластеризации и тематического моделирования:

- **Silhouette Score**: 0.5397239923477173
- **Coherence (c_v)**: 0.8210221614838007
- **Coherence (c_npmi)**: 0.19905209113081204
- **Coherence (u_mass)**: -2.8766630095168915
- **Coherence (c_uci)**: 0.8559725662682557

- **Topic Diversity**: 0.9411764705882353

Также было представлено распределение топиков в двумерном пространстве:



На графике видно, что топики получились достаточно разнообразными. Например, в нижней части визуализации выделились топики (0, 33, 17), связанные с использованием видеогенерации в медицине.

Далее были выделены топики, непосредственно касающиеся видеогенерации.

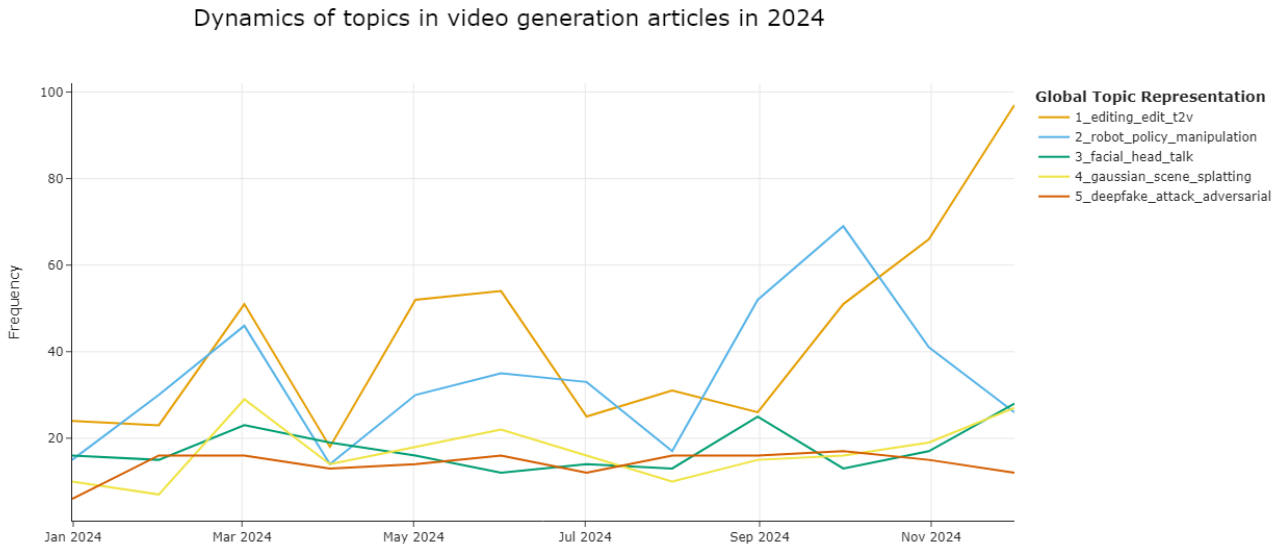
Опираясь на базовые знания в области видеогенерации (хотя основная специализация — NLP) и результаты поиска в Google, был сформирован набор тем для дальнейшего анализа. Ниже приведены основные направления и несколько дополнительных тем: Основные темы:

- 1_editing_edit_t2v: генерация и редактирование видео по текстовым описаниям.
- 2_robot_policy_manipulation: генерация видео используется для демонстрации или обучения манипуляционным действиям в робототехнике.
- 3_facial_head_talk: синтез говорящих голов, где по изображению и аудио создаётся реалистичная анимация лица.
- 4_gaussian_scene_splatting: рендеринг динамических сцен с использованием 4D Gaussian Splatting,
- 5_deepfake_attack_adversarial: генерация дипфейков.

Дополнительные темы:

- 18_character_synthesis_body: синтез персонажей (возможное применение для анимации или создания персонажей в видео)
- 26_emotion_expression_facial: генерация эмоциональных выражений на лицах

Основные темы были выделены также с учётом размера топиков (они достаточно крупные), а дополнительные темы выбраны из интереса.



Из представленного графика, видно, что

- К концу 2024 года лидируют два топика 2_robot_policy_manipulation (всплеск в ноябре—декабре) и 1_editing_edit_t2v (стабильный рост в последнем квартале).
- 5_deepfake_attack_adversarial также показывает устойчивый подъём, но не столь резкий.
- 3_facial_head_talk держится на среднем уровне без экстремальных колебаний.
- 4_gaussian_scene_splatting остаётся относительно нишевой темой с единичными всплесками интереса.

Также я решил провести регрессионный анализ для выявления статистически значимого тренда. Результаты анализа за 2024 год продемонстрировали следующие показатели:

Таблица 1: Regression results by topics (with HAC robust errors)

	1_editing_edit_t2v	2_robot_policy_manipulation	3_facial_head_talk	4_gaussian_scene_splatting	5_deepfake_attack_adversarial
	(1)	(2)	(3)	(4)	(5)
const	28.400***	35.486***	17.457***	17.171***	18.486***
	(9.288)	(6.951)	(2.157)	(3.375)	(2.257)
month_num	2.705	0.232	0.250	0.186	-0.452
	(2.041)	(1.239)	(0.339)	(0.459)	(0.308)
Observations	11	11	11	11	11
R^2	0.163	0.002	0.024	0.008	0.091
Adjusted R^2	0.070	-0.108	-0.085	-0.102	-0.010
Residual Std. Error	22.485 (df=9)	17.462 (df=9)	5.889 (df=9)	7.383 (df=9)	5.227 (df=9)
F Statistic	1.757 (df=1; 9)	0.035 (df=1; 9)	0.543 (df=1; 9)	0.164 (df=1; 9)	2.150 (df=1; 9)
Note:					*p<0.1; **p<0.05; ***p<0.01

Из представленной таблицы видно, что для всех тем статистически значимых трендов выявлено не было.

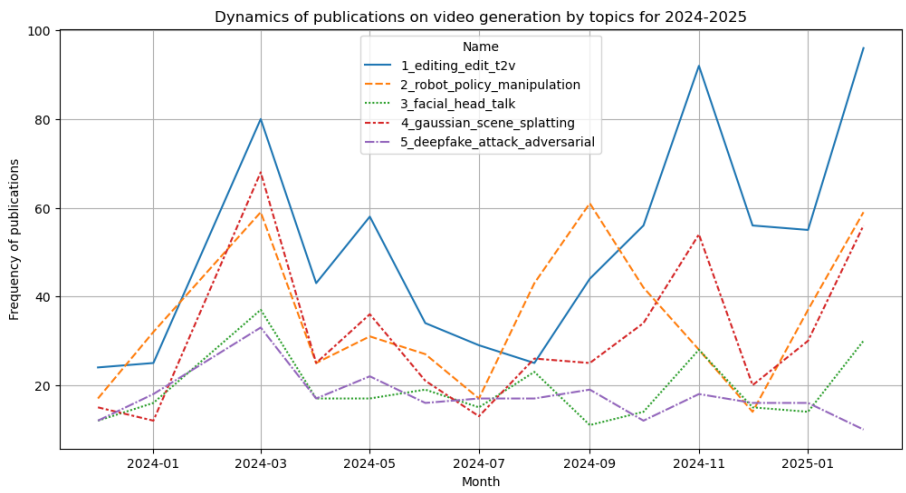
Чтобы более полно проследить динамику, я расширил анализ, добавив данные за 2025 год. Полученные результаты представлены ниже:

Таблица 2: Regression results by topics (with HAC robust errors)

	1_editing_edit_t2v	2_robot_policy_manipulation	3_facial_head_talk	4_gaussian_scene_splatting	5_deepfake_attack_adversarial
	(1)	(2)	(3)	(4)	(5)
const	56.053***	44.885***	22.784***	36.996***	19.673***
	(11.179)	(4.535)	(3.000)	(6.732)	(3.451)
month_num	-0.744	-1.499**	-0.560	-0.912	-0.356
	(1.616)	(0.725)	(0.366)	(0.851)	(0.355)
Observations	14	14	14	14	14
R^2	0.015	0.140	0.085	0.044	0.067
Adjusted R^2	-0.067	0.068	0.008	-0.035	-0.011
Residual Std. Error	25.026 (df=12)	15.315 (df=12)	7.582 (df=12)	17.435 (df=12)	5.488 (df=12)
F Statistic	0.212 (df=1; 12)	4.269* (df=1; 12)	2.346 (df=1; 12)	1.147 (df=1; 12)	1.007 (df=1; 12)
Note:					*p<0.1; **p<0.05; ***p<0.01

Из этой таблицы видно, что добавление данных за 2025 год выявило статистически значимое снижение интереса к топику 2_robot_policy_manipulation ($p<0.05$). Однако интерпретировать эти результаты стоит с осторожностью, поскольку используется простая линейная регрессия с лагами и робастными стандартными ошибками, а количество наблюдений остаётся невысоким.

Для более наглядного анализа обратим внимание на график:



На графике отчетливо прослеживается рост публикаций по теме генерации видео по текстовым описаниям, а также динамика по остальным топикам, за исключением генерация дипфейков. Визуальный анализ в данном случае предоставляет больше информации, так как число наблюдений для регрессионного анализа недостаточно для окончательных выводов.

Подводя итоги обзора, можно сделать следующие выводы. Динамика публикаций по видеогенерации на `arXiv` за 2024 год демонстрирует общий положительный тренд, особенно в темах, связанных с генерацией видео по текстовым описаниям и обучением манипуляционным действиям в робототехнике. Хотя регрессионный анализ с данными 2025 года выявил статистически значимое снижение интереса к топику 2_robot_policy_manipulation, визуальный анализ не подтверждает эти результаты. Это может быть связано с ограниченным числом наблюдений, что указывает на необходимость дальнейшего исследования с расширенной выборкой для более точного понимания динамики развития направлений в видеогенерации.

3 Сбор данных

В этом разделе продемонстрирован процесс формирования базы данных с использованием библиотеки `arXiv` – удобной обёртки для API-запросов. Изучив синонимичные запросы, охватывающие различные аспекты генерации видео, я выделил следующий перечень ключевых тем:

- video generation
- text-to-video
- video synthesis
- generative video
- video diffusion
- long video generation
- video transformer
- motion synthesis
- spatiotemporal generation
- video autoregressive
- video GAN

Концептуально важным является использование запроса вида: ”all:video AND all:generation”, который обеспечивает сбор статей, содержащих оба термина, в отличие от альтернативного запроса ”all: video generation”, выдающего значительно меньше результатов (около 700 статей). Такой подход позволяет охватить большое количество публикаций, но, с другой стороны, приводит к зашумлению данных. Для минимизации шума применяется метод

HDBSCAN, способный выделять шумовой кластер. Хотя данная стратегия не является идеальной, в условиях широкого охвата данных её использование оправдано и эффективно.

В результате выполнения кода была сформирована база из 8870 статей, из которых 6915 относятся к 2024 году. Извлечены следующие ключевые метаданные:

- entry_id — уникальный идентификатор статьи,
- title — название статьи,
- abstract — краткое содержание статьи,
- published — дата публикации в формате ГГГГ-ММ-ДД,
- primary_category — основная категория статьи на arXiv.org.

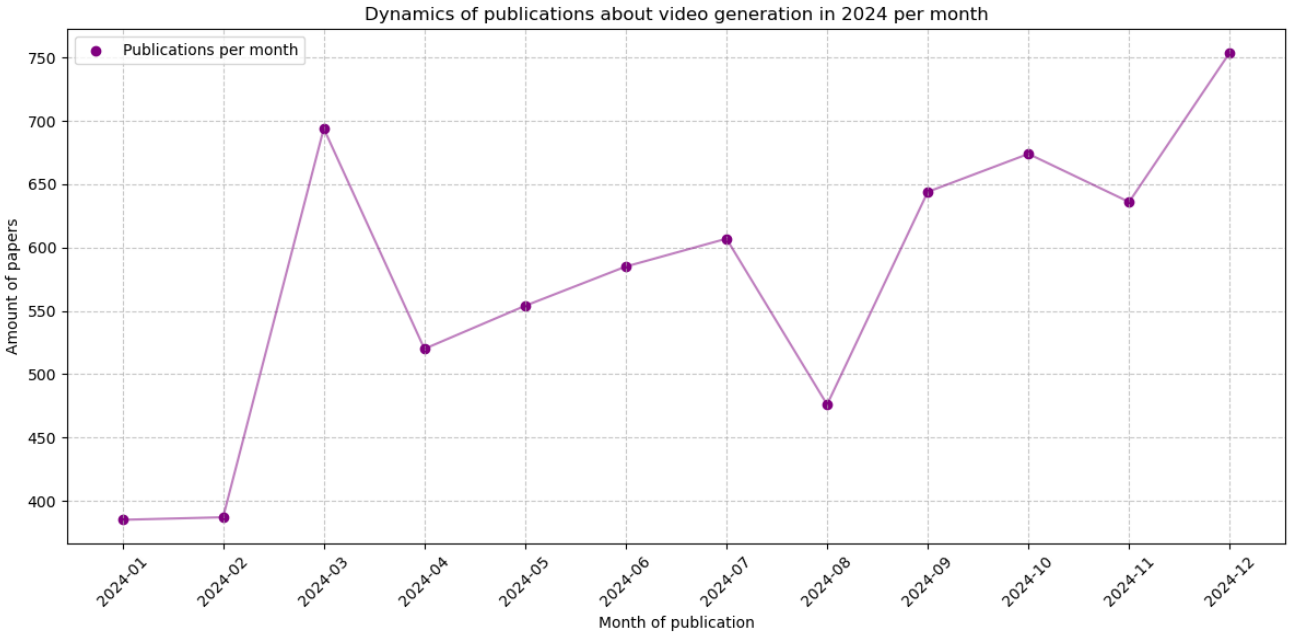
В перспективе возможно внедрение асинхронного кода для ускорения процесса сбора, хотя текущая скорость отклика API удовлетворительна.Итоговые результаты сохранены в папку data:

1. arxiv_video_generation_papers_2024_2025.csv — файл с метаданными статей по видеогенерации за 2024 год и начало 2025 года,
2. video_generation_2024.csv — файл с метаданными статей по видеогенерации за 2024 год.

4 Обзор собранных данных

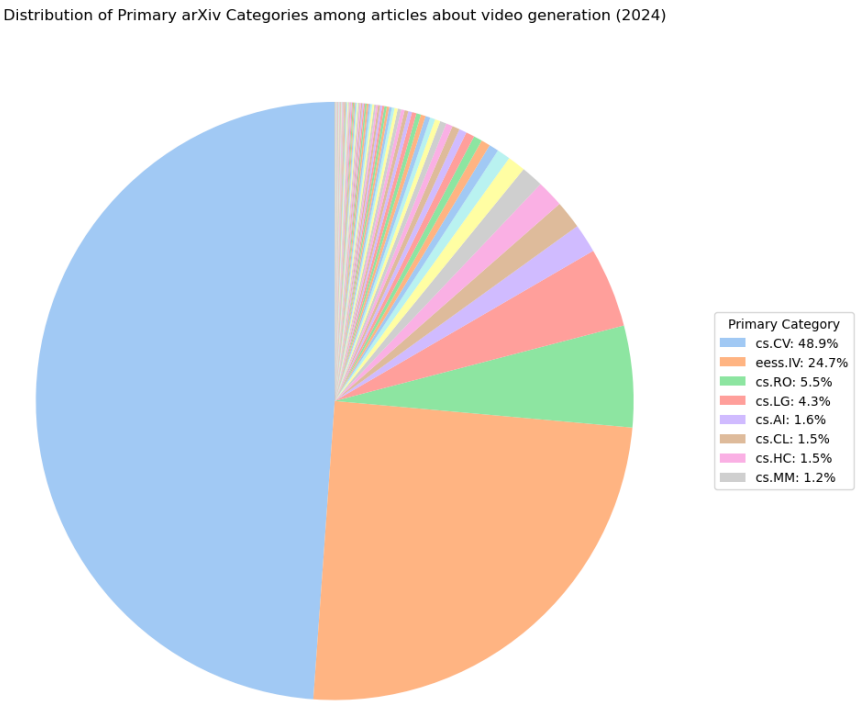
Был получен набор данных для анализа, и теперь предлагается перейти к детальному рассмотрению, чтобы оценить его "адекватность". В данном блоке хотим убедиться, что применение запроса ”all:video AND all:generation” не привело к чрезмерному шуму, а также проверить представленность статей на протяжении всего 2024 года.

В первую очередь необходимо оценить динамику публикаций, чтобы удостовериться, что данные охватывают каждый месяц. На приведённом графике видно, что за каждый месяц имеется более 400 статей, а также прослеживается общий положительный тренд в появлении публикаций по видеогенерации. Однако окончательные выводы делать пока рано, поскольку данные могут содержать шум.



Кроме того, каждая статья на `arXiv` сопровождается меткой `primary_category`, которая определяет её принадлежность к определённой области. Более подробную информацию о метках можно найти на официальном сайте arXiv.org.

Ниже представлен barchart, иллюстрирующий распределение статей по основным категориям. Из диаграммы видно, что топ-2 категории — `cs.CV` (Computer Vision and Pattern Recognition) и `eess.IV` (Image and Video Processing). Это свидетельствует об адекватности подхода к сбору данных, несмотря на наличие множества аутлайеров в виде небольших категорий. Использование HDBSCAN для устранения шума является оправданным; в дальнейшем рекомендуется обратить внимание на подозрительные кластеры и соотнести их с метками `primary_category`.



5 Обучение модели и метрики качества

5.1 Выбор модели и обоснование

Для выявления тематических направлений и отслеживания динамики развития научных публикаций в области видеогенерации была выбрана модель BERTopic. Данный подход оказался оптимальным по нескольким причинам:

- Использование контекстуальных эмбеддингов на основе современных языковых моделей позволяет учитывать специфику научной лексики.
- Кластеризация с применением алгоритма `HDBSCAN` обеспечивает устойчивое выделение тематических кластеров и эффективное отделение шумовых наблюдений.
- Названия тем формируются автоматически с помощью `TF-IDF`, где для каждой темы отбираются три наиболее значимых ключевых слова, что упрощает интерпретацию топиков.
- Модель поддерживает временной анализ, позволяя отслеживать эволюцию тем по месяцам.

5.2 Архитектура тематического моделирования

Процесс тематического моделирования включал несколько этапов:

1. **Предобработка данных.** Тексты очищались от гиперссылок и специальных символов, а также приводились к нормальной форме посредством лемматизации. При построении эмбеддингов стоп-слова не удалялись, чтобы сохранить контекст.
2. **Формирование входных текстов.** Для каждого документа использовалась конкатенация заголовка статьи и её аннотации, что обеспечивало более полное представление содержания.
3. **Получение эмбеддингов.** В качестве энкодера применялась модель `all-MiniLM-L6-v2`, которая обеспечивает оптимальный баланс между качеством представлений и вычислительной эффективностью.
4. **Снижение размерности.** Для снижения размерности нелинейных эмбеддингов использовался метод `UMAP` с параметрами:
   - `n_components=15`,
   - `n_neighbors=15`,
   - `min_dist=0.05`,
   - `metric='cosine'`,
   - `random_state=42`.

   Данный метод позволяет сохранить топологическую структуру данных, что является критически важным для последующей кластеризации.

5. **Кластеризация.** Для группировки документов был применён алгоритм `HDBSCAN`, позволяющий выделять плотные кластеры и отделять шумовые наблюдения.

6. **Извлечение ключевых слов.** Для интерпретации полученных кластеров использовался `CountVectorizer` с параметрами `max_df=0.8` и `min_df=2`. Список стоп-слов был дополнен словами, характерными для научных текстов. Для каждой темы автоматически выбирались три наиболее значимых термина по метрике `TF-IDF`.

5.3 Технические особенности

Для ускорения вычислений использовалась проприетарная библиотека NVIDIA `cuml`, позволяющая выполнять часть операций на GPU. Предобработка текстов была реализована в многопоточном режиме с помощью библиотеки `pandarallel`. Несмотря на то, что современные энкодеры зачастую не требуют обширного препроцессинга, удаление ссылок и лишних символов оказалось необходимым для снижения уровня шума.

5.4 Оптимизация гиперпараметров и метрики качества

В процессе обучения параметры модели подбирались с целью максимизации **Silhouette Score** и **c_v**, так как, согласно исследованиям (см. Berksudan, 2018), метрика `c_v` лучше коррелирует с человеческой оценкой тематической связности по сравнению с альтернативными показателями. Используемые метрики для оценки качества модели включают:

- **Silhouette Score** (от -1 до 1) — отражает степень разделения кластеров; значения выше 0.5 считаются хорошими.
- **c_v** (от 0 до 1) — мера семантической связности тем; значения выше 0.5 свидетельствуют о приемлемой когерентности.
- **c_npmi** (от -1 до 1) — нормированная точечная взаимная информация, где положительные значения указывают на значимую связь между словами в теме.
- **u_mass** (от $-\infty$ до 0) — когерентность, основанная на совместной встречаемости слов (ближе к 0 — лучше).
- **c_uci** — мера когерентности на основе точечной взаимной информации (чем выше, тем лучше).
- **Topic Diversity** (от 0 до 1) — доля уникальных слов в топ-N списках для тем; значения, приближающиеся к 1, указывают на разнообразие тем.

После подбора гиперпараметров итоговая модель продемонстрировала следующие показатели:

- **Silhouette Score**: 0.5397239923477173
- **Coherence (c_v)**: 0.8210221614838007
- **Coherence (c_npmi)**: 0.19905209113081204

- **Coherence (u_mass)**: -2.8766630095168915
- **Coherence (c_uci)**: 0.8559725662682557
- **Topic Diversity**: 0.9411764705882353

При кластеризации было выделено 35 топиков, а 2132 документа были отнесены к шумовому кластеру. Детальный анализ шумового кластера позволит в дальнейшем уточнить границы тематической структуры.

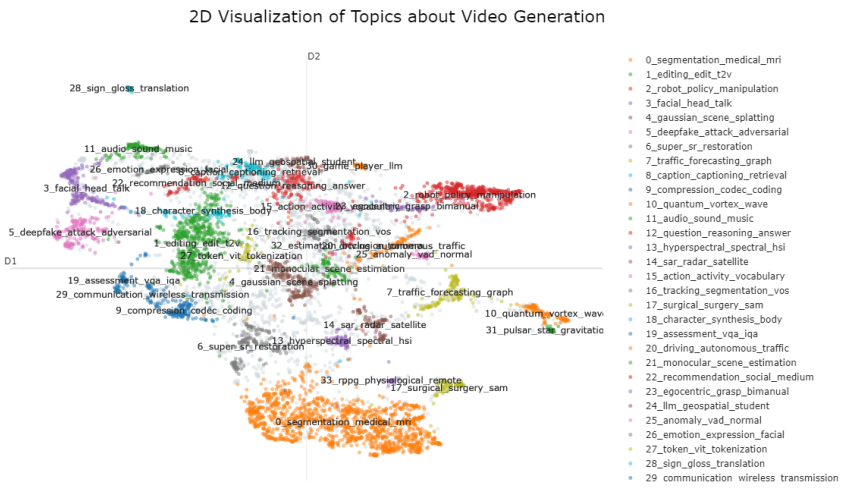
Итоговая модель была сохранена в папку `bertopic_model`, а предобработанный датасет — в файл `video_generation_2024_bertopic.csv`.

6 Обзор полученных тем и визуальный анализ

В данном блоке использована обученная модель `bertopic_model` из ноутбука `3_topic_modeling.ipynb`. Основная цель данного этапа заключалась в следующем:

- Визуализировать все выделенные топики в двухмерном пространстве.
- Оценить адекватность топиков в контексте видеогенерации и выделить наиболее релевантные темы.
- Построить динамику появления топиков по месяцам.

Для визуализации топиков был использован метод отображения в двумерном пространстве. Следует отметить, что данная визуализация предназначена исключительно для иллюстративных целей, поскольку проекция в двух измерения «съедает» значительную часть информации о структуре топиков, и расстояния между ними интерпретировать напрямую нецелесообразно.



Анализ графика демонстрирует, что топики получились достаточно разнообразными. В частности, в нижней части визуализации отчетливо выделяются топики с идентификаторами 0, 33 и 17, которые связаны с применением видеогенерации в медицине.

Опираясь на базовые знания в области видеогенерации и результаты поиска в Google, сформирован набор тем, заслуживающих дальнейшего анализа. Темы были разделены на две группы: основные и дополнительные.

6.1 Основные темы

- `1_editing_edit_t2v`: генерация и редактирование видео по текстовым описаниям.
- `2_robot_policy_manipulation`: генерация видео для демонстрации или обучения манипуляционным действиям в робототехнике.
- `3_facial_head_talk`: синтез говорящих голов, при котором по изображению и аудио создаётся реалистичная анимация лица.
- `4_gaussian_scene_splatting`: рендеринг динамических сцен с использованием 4D Gaussian Splatting.
- `5_deepfake_attack_adversarial`: генерация дипфейков.

6.2 Дополнительные темы

- `18_character_synthesis_body`: синтез персонажей, потенциально применимый для анимации или создания видеоперсонажей.
- `26_emotion_expression_facial`: генерация эмоциональных выражений на лицах.

Классификация тем проведена с учётом размеров топиков, что позволило выделить более крупные, релевантные направления, в то время как дополнительные темы были включены в анализ из интереса.

Из представленной визуализации видно, что:

- К концу 2024 года доминируют два топика: `2_robot_policy_manipulation` (отмечен всплеском в ноябре—декабре) и `1_editing_edit_t2v` (демонстрирует стабильный рост в последнем квартале).
- Топик `5_deepfake_attack_adversarial` показывает устойчивый, хотя менее резкий, подъём.
- Топик `3_facial_head_talk` сохраняет средний уровень активности без резких колебаний.
- Топик `4_gaussian_scene_splatting` остаётся нишевым, с единичными всплесками интереса.

Динамика топиков по месяцам была также сохранена в файл `topics_over_time.csv`

При этом визуального анализа недостаточно для полноценной интерпретации динамики трендов, поэтому было принято решение о проведении прилиминарного статистического анализа.

7 Статистический анализ тренда

В рамках анализа динамики публикаций по тематическим направлениям, полученным в результате тематического моделирования, агрегировались данные по месяцам (всего 12 наблюдений для каждого топика). Выбор агрегации по месяцам был обусловлен тем, что недельная агрегация привела бы к избыточной шумности данных, что затруднило бы анализ тренда.

Для оценки временной динамики была использована линейная регрессия следующего вида:

$$y_i = \beta_0 + \beta_1 \times \text{month\_num}_i + \epsilon_1$$

где:

- y_i — количество публикаций по конкретному топику в i-м месяце;
- β_0 — константа (интерсепт), отражающая базовый уровень публикаций;
- β_1 — коэффициент, характеризующий изменение количества публикаций при изменении номера месяца (month_num);
- ε_i — случайная ошибка для i-го наблюдения.

Для корректного учета возможной нестатичности дисперсии во временном ряду были использованы HAC-робастные стандартные ошибки.

Результаты анализа за 2024 год не выявили статистически значимых трендов по всем темам.

Таблица 3: Regression results by topics (with HAC robust errors)

	1_editing_edit_t2v	2_robot_policy_manipulation	3_facial_head_talk	4_gaussian_scene_splatting	5_deepfake_attack_adversarial
	(1)	(2)	(3)	(4)	(5)
const	28.400***	35.486***	17.457***	17.171***	18.486***
	(9.288)	(6.951)	(2.157)	(3.375)	(2.257)
month_num	2.705	0.232	0.250	0.186	-0.452
	(2.041)	(1.239)	(0.339)	(0.459)	(0.308)
Observations	11	11	11	11	11
R^2	0.163	0.002	0.024	0.008	0.091
Adjusted R^2	0.070	-0.108	-0.085	-0.102	-0.010
Residual Std. Error	22.485 (df=9)	17.462 (df=9)	5.889 (df=9)	7.383 (df=9)	5.227 (df=9)
F Statistic	1.757 (df=1; 9)	0.035 (df=1; 9)	0.543 (df=1; 9)	0.164 (df=1; 9)	2.150 (df=1; 9)
Note:					*p<0.1; **p<0.05; ***p<0.01

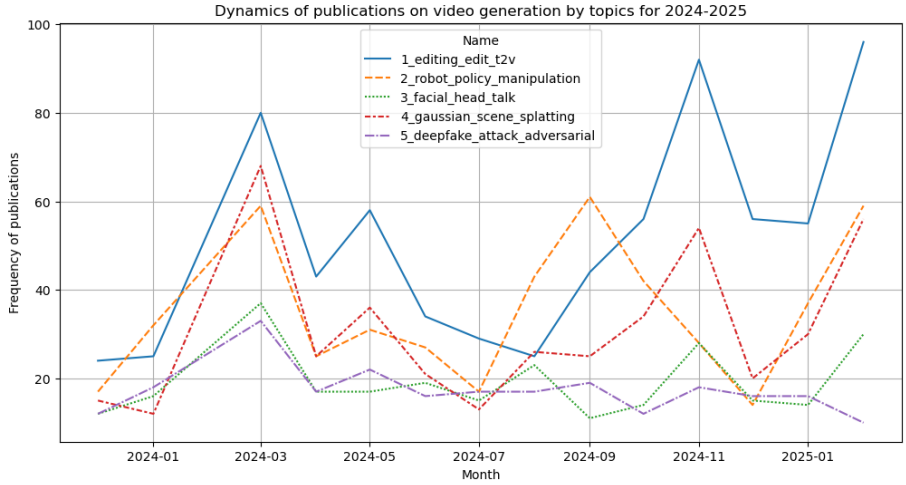
Однако при расширении анализа с включением данных за 2025 год наблюдалось статистически значимое снижение интереса к топику `2_robot_policy_manipulation` (p<0.05).

Таблица 4: Regression results by topics (with HAC robust errors)

	1_editing_edit_t2v	2_robot_policy_manipulation	3_facial_head_talk	4_gaussian_scene_splatting	5_deepfake_attack_adversarial
	(1)	(2)	(3)	(4)	(5)
const	56.053***	44.885***	22.784***	36.996***	19.673***
	(11.179)	(4.535)	(3.000)	(6.732)	(3.451)
month_num	-0.744	-1.499**	-0.560	-0.912	-0.356
	(1.616)	(0.725)	(0.366)	(0.851)	(0.355)
Observations	14	14	14	14	14
R^2	0.015	0.140	0.085	0.044	0.067
Adjusted R^2	-0.067	0.068	0.008	-0.035	-0.011
Residual Std. Error	25.026 (df=12)	15.315 (df=12)	7.582 (df=12)	17.435 (df=12)	5.488 (df=12)
F Statistic	0.212 (df=1; 12)	4.269* (df=1; 12)	2.346 (df=1; 12)	1.147 (df=1; 12)	1.007 (df=1; 12)
Note:					*p<0.1; **p<0.05; ***p<0.01

Следует отметить, что интерпретация этих результатов требует осторожности ввиду использования простой линейной регрессии с лагами и ограниченного числа наблюдений.

Для более наглядной интерпретации динамики публикаций представлен график:



Из графика видно, что отчетливо прослеживается рост активности по теме генерации видео по текстовым описаниям, а также различная динамика по остальным топикам, за исключением генерации дипфейков. Визуальный анализ дополнительно иллюстрирует тенденции, учитывая, что число наблюдений для регрессионного анализа недостаточно для окончательных выводов.

8 Заключение

В рамках данного исследования проведён анализ тематики статей по видеогенерации, опубликованных на `arXiv` в 2024 году.

Этот анализ включал сбор и предобработку данных, построение тематической модели для выделения основных направлений и последующий регрессионный анализ динамики выявленных тем. Анализ позволил выделить несколько ключевых тематических направлений; наиболее актуальными среди них оказались темы, связанные с диффузионными моделями и генерацией видео на основе текстовых описаний, которые продемонстрировали значительный рост числа публикаций в течение года, указывающий на положительный тренд интереса исследовательского сообщества.

Таким образом, полученные результаты отражают текущую структуру исследований в области генерации видео и свидетельствуют о наличии активно развивающихся направлений.

Тем не менее, интерпретация результатов ограничена рядом факторов. Во-первых, анализ охватывал только один год наблюдений, что затрудняет выявление долгосрочных тенденций развития тематики. Во-вторых, среди выделенных тем присутствовал «шумовой» кластер, состоящий из разнородных или малоинформативных документов, что указывает на наличие шума в данных и ограничения применённого подхода к тематическому моделированию. В-третьих, регрессионная оценка трендов основана на ограниченном числе временных точек (месяцев), поэтому статистическая значимость выявленных тенденций невысока и их интерпретация может быть неоднозначной. Перспективы дальнейшей работы включают улучшение тематической модели — например, более тонкую настройку параметров или использование более совершенных алгоритмов для снижения влияния шума и более чёткого разделения тематик.

Также представляет интерес детальный анализ документов из «шумового» кластера с целью выяснить природу этого шума и, возможно, скорректировать критерии отбора данных.

Кроме того, расширение временного охвата выборки (например, анализ данных за несколько лет) и применение более строгих методов анализа трендов (с учётом статистической значимости изменений) позволят получить более надёжные выводы о динамике развития тематических направлений в генерации видео.