Txt2Vid-Web: Web-based, Text-to-Video, Video Conferencing Pipeline

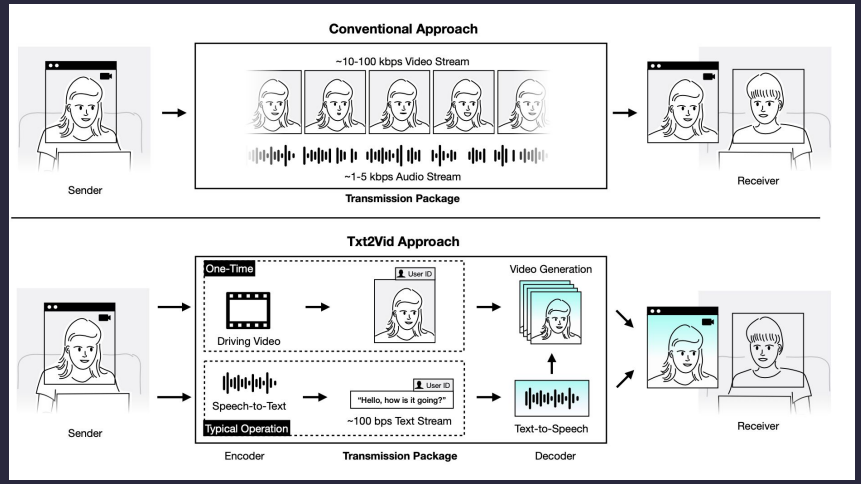
Arjun Barrett, Laura Gomezjurado, Shuvam Mukherjee, Arz Bshara, Sahasrajit Sarmasarkar, Pulkit Tandon and Tsachy Weissman

INTRODUCTION

Video Conferencing has become increasingly **relevant**. However, existing platforms require **several Mbps of bandwidth**, reducing **multimedia access** especially in underserved regions with poor internet connectivity.

Txt2Vid

Video compression pipeline. Synthesizes video by lip-syncing existing footage to **text-to-speech** results from **deep-fake** voice clones. **Achieved 1000X compression advantage.**



Tandon, P. (2021, June 26). https://arxiv.org/abs/2106.14014

Architecture

We implement the Txt2Vid pipelines proposed with a voice cloning model, lip-syncing model, on the browser application stack – making it more accessible and portable.

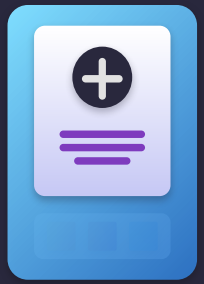
We enabled acceleration at the decoder via WebGL,by implementing a new WebGL shader for ConvTranspose in ONNX Runtime Web.

We verified our platform via subjective study and show that at similar quality of experience our platform requires 100-500X less bandwidth than standard compressors.

IMPLEMENTATION

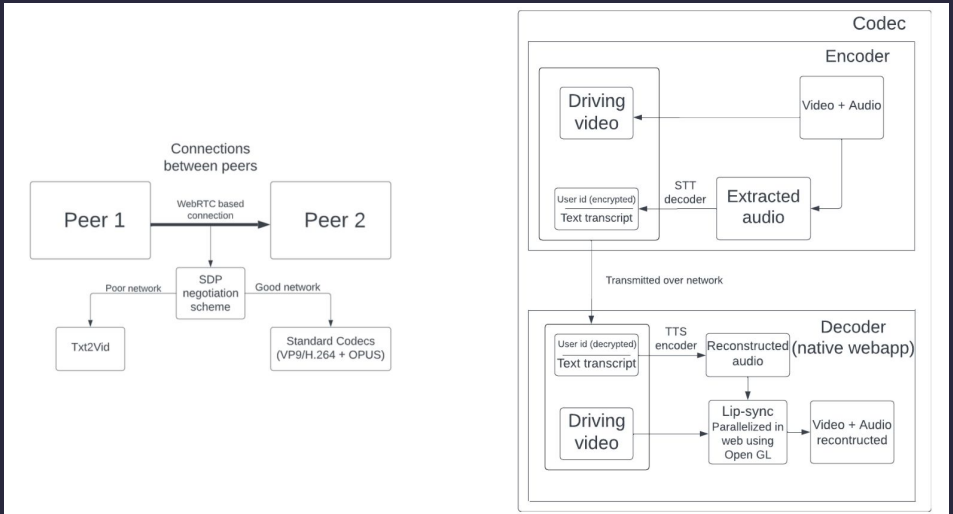
Our implementation is available as a **web application.**

Employed **ONNX Runtime Web**, an efficient neural network inference engine, along with **WebGL-based GPU acceleration** and fine-tuned preprocessing logic.



01

Block diagram of the web-conferencing platform



(a) SDP negotiation scheme (b) Txt2Vid-Web Pipeline

Seemless switching between Txt2Vid and traditional codecs using WebRTC

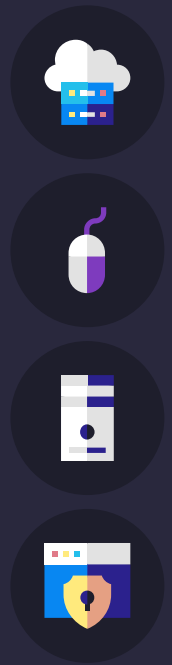
Modifications

To TTS

Used Resemble.ai as TTS generator. Protected the security of credentials encrypting each API key. Each peer can perform TTS generation for the original user's voice and satisfy audio generation for Txt2Vid decoding without sharing secret keys.

To pipeline

Employed Progressive Web App. When website code and pre-trained Wav2Lip model are downloaded, the site saves the files to the computer, will not require re-downloading.



To Lip-syncing

Ported the Wav2Lip neural network from PyTorch to ONNX to enable usage within users' web browsers. Switched to a compute backend based on WebGL. Created a novel shader for ConvTranspose and contributed it back to the ONNX Runtime Web neural network inference engine as open-source software via a pull request.

EVALUATION

We compared **user-perceived quality-of-experience** (QoE) of Txt2Vid with that obtained **by traditional codecs** (H.264 and VP9 for video and OPUS for audio) at various bitrates

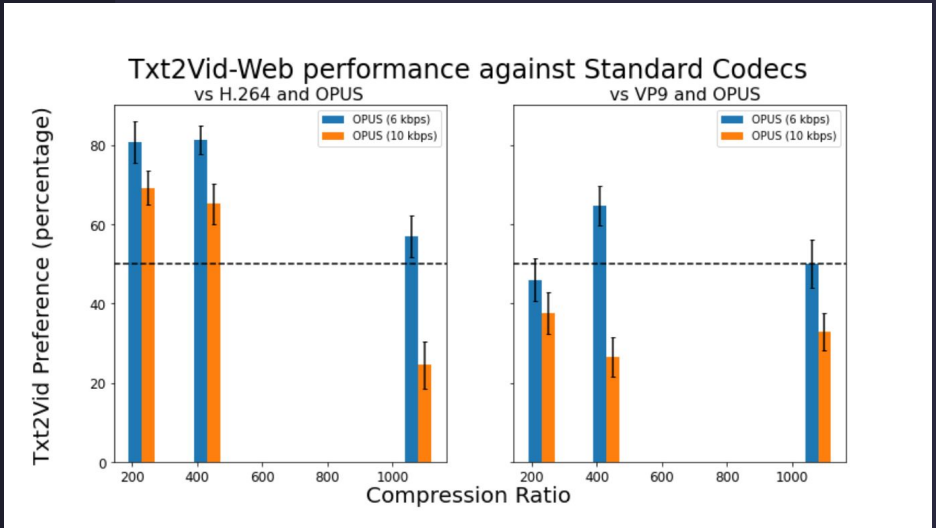
Dataset

Table 1: Difference between Txt2Vid and Baseline Codec Ratings in the range of 1-5. Positive values imply Txt2Vid is preferred over baseline codec.

	H.264 (15 kbps)	H.264 (35 kbps)	H.264 (100 kbps)	VP9 (15 kbps)	VP9 (35 kbps)	VP9 (100 kbps)
OPUS (6 kbps)	1.29	0.95	0.16	0.16	0.58	0.40
OPUS (10 kbps)	0.83	0.60	-0.74	-0.38	-0.56	-0.68

5 ~1 min talking-head videos of 5 people with each person speaking. > **Baseline codec dataset** using traditional video and audio codecs, bitrates ranging from 6k to 100k. > A total of 5 × 2 × 6 = 60 pairs of videos which were evaluated. > Asked human subjects to watch the pairs, rate each video, and choose which one they preferred.

Evaluation Results



Txt2Vid-Web either outperforms or performs equivalently to standard codecs based on QoE magnitude.

To obtain a similar QoE under traditional codecs, we require at least 100X less bandwidth.

Against the most common codec combination (OPUS-H.264), Txt2Vid-Web is 60% preferred with 0.25% bandwidth

The obtained gains are lower, but still substantial, for the more modern codec pair of OPUS-VP9.

03

APPLICATIONS

We envision our web-based video conferencing platform can open up many new potential applications, **enabling audiovisual communications** even in contexts where traditional codecs and techniques fail.

Potential Use Cases

REMOTE VIDEO FOR RURAL AREAS

Telemedicine
Receive medical care despite geographical barriers

Education
Enabling online learning with live or pre-recorded text lectures

POOR CONNECTIVITY AUTO-FALLBACK

Use high-quality video codec by default for video call

Fall back to Txt2Vid when connection bandwidth drops

LARGE-SCALE VIDEO CONFERENCING

As consumer devices become more powerful, can receive high quality video feeds from thousands of participants at once via Txt2Vid-encoded video streams

Conclusions

We find that Txt2Vid-Web requires orders of magnitude lower bandwidth for comparative QoE by focusing on keeping the perceptual audio-video features intact and disregarding the low-level details such as video pixel or audio sample fidelity.

As the first implementation of the Txt2Vid framework proven to work on consumer devices, Txt2Vid-Web unlocks a wide variety of real-world applications and offers solutions to many of the practical challenges machine-learning based compression techniques face.

FURTHER RESOURCES

Txt2Vid Paper:
IEEE Journal on Selected Areas in Communication, 2022

Txt2Vid-Web Repository:
https://github.com/tpulkit/txt2vid_browser

THANK YOU