Computer vision and multimodal AI

How models see and listen

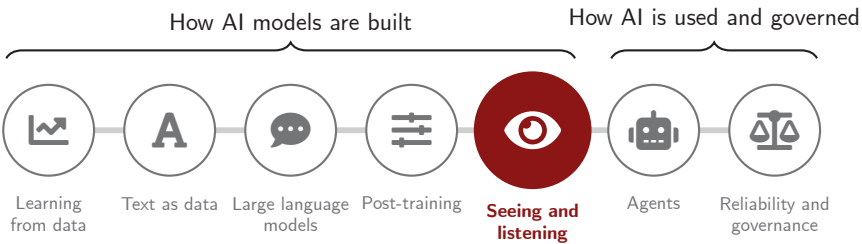
POLI 158: AI & Politics · Session 13

Yiqing Xu

TODAY'S QUESTION

A model looks at a photo of a candidate and guesses the candidate's politics. What is it really looking at?

The seven Wednesday sessions



Sessions are numbered across Mondays and Wednesdays: today is Session 13.

Today in five parts



Where computers see and listen

Everyday life


searching your photos


live captions


talking to a chatbot

Research


finding protest photos


coding debate videos


measuring voices in parliament

Scholz et al. (2025); Mestre and Ryan (2026); Rittmann (2024)

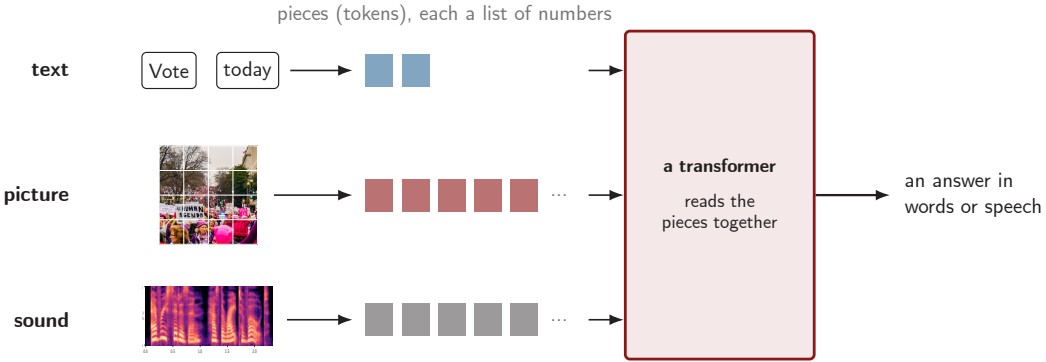
Politics and policy


describing candidates

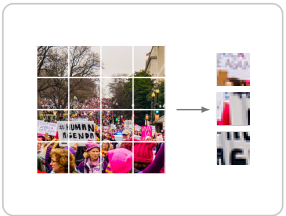

face recognition


transcribing hearings

Text, pictures, and sound become pieces

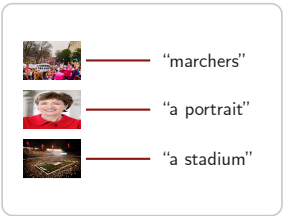


Three concepts in today's video



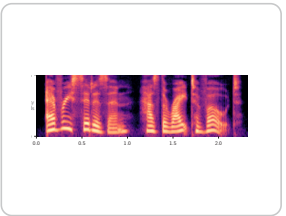
Patches

a photo becomes tokens



A shared map

photos and words in one space



Sound and video

voices and videos become tokens

Video part 1: concepts 1 and 2. Part 2: concept 3.

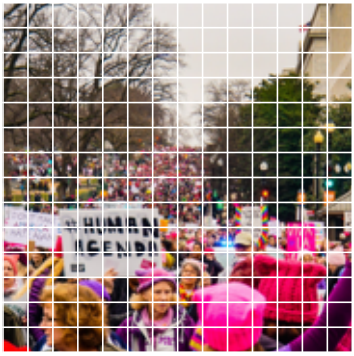


How models see and listen, part 1

Video

Key ideas: patches and a shared map

A photo as patches



- ViT (2020): 224 × 224 pixels, in 16 × 16 squares.
- 14 × 14 = 196 patches, one token each.



one row of the photo, in reading order

Source: Dosovitskiy et al. (2021), pp. 1, 3, 5, 13 (our arithmetic). Photo: Ted Eytan, Women's March, Washington, 2017, CC BY-SA 2.0 (creativecommons.org/licenses/by-sa/2.0); cropped, resized, grid added.

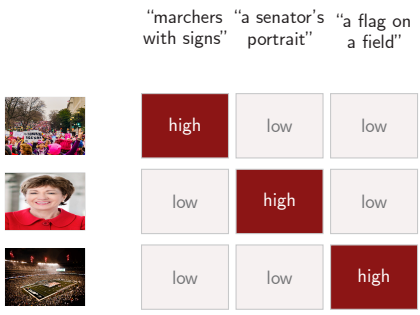
What a photo costs in tokens

A photo of 1,000 × 1,000 pixels	
Claude: 36 × 36 squares	1,296 tokens
About 970 English words	about the same

- Each square, 28 pixels wide, is a token.
- The companies bill by the token.
- Big photos are shrunk first: small print blurs.

Source: Anthropic, "Vision" documentation, checked October 9, 2026. Words: 100 tokens are about 75 English words (OpenAI); our arithmetic.

CLIP: matching photos and captions



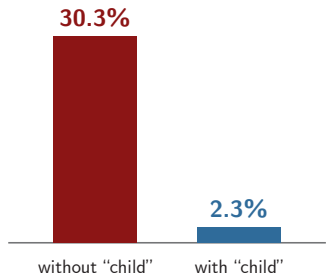
- Photo encoder: photo → numbers.
- Text encoder: text → numbers.
- Training pulls each photo toward its own caption on a map.
- 400 million photo–caption pairs from the internet.
- Then ask: does "a photo of a protest" fit?

Illustration: match scores. In training, each grid had 32,768 photos and 32,768 captions.

Source: Radford et al. (2021), "Learning Transferable Visual Models From Natural Language Supervision," pp. 1, 4–5, and 7.

Who writes the labels?

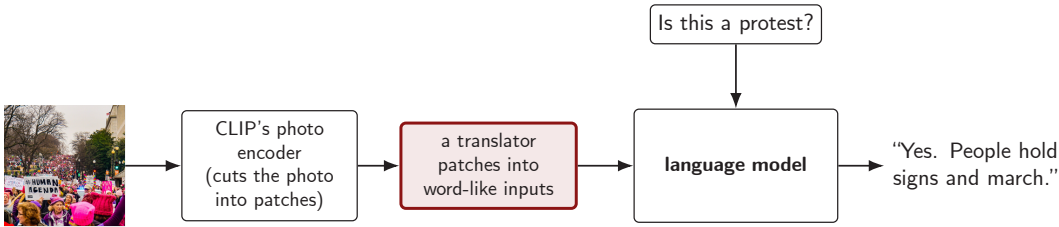
- CLIP's authors: 10,000 face photos, with labels like "thief" and "animal."
- Adding the label "child": 30.3% → 2.3% (ages 0 to 2).
- The label list changes the result.



Ages 0 to 2: share labeled as a crime or non-human class.

Source: Radford et al. (2021), section 7.1, p. 22, and Table 7.

From CLIP to a chatbot that sees



- LLaVA (2023): the translator is one table of learned weights.

Source: Liu et al. (2023), "Visual Instruction Tuning," pp. 3–5. The answer is an illustration.

Activity: what the model sees and hears

Activity, part 1: fool the camera

Groups of three or four. One phone per group, with a chatbot that can see: ChatGPT, Gemini, or Claude.

1. Point the camera at something on your desk.
2. Ask a question with an answer you can check.
3. Goal: a confident wrong answer. You have 6 minutes.
4. One point for each one. "I'm not sure" scores nothing.

Share: your best fool, and why you think it was hard.

Things to try

How many pens are in this pile?

What time does this watch show?

Read this handwriting, word for word.

What is this? (an object upside down, or half hidden)

Rule: objects only. No faces, and no screens with your own information: the photos go to the chatbot's company.

What CLIP said about the four photos

Four photos of crowds, given to CLIP with eight labels to choose from.

1



March on Washington, 1963

"campaign rally" (0.94)

wrong

2



suffragists' picket, 1917

"protest" (0.52)

right, narrowly

3



Sanders rally, 2020

"campaign rally" (0.86)

right

4



vigil and protest, 2022

"vigil" (1.00)

partly

- "Protest" was on the list. The 1963 march still got "campaign rally."
- Scores sum to 1 across the labels: a preference, not a chance of being right.

CLIP ViT-B/16 (the model in a public demo on Hugging Face), run on this laptop, October 10, 2026. Labels: backup page.

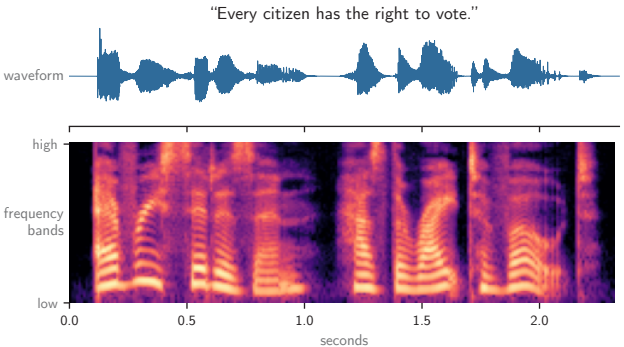


How models see and listen, part 2

Video

Key ideas: sound and video

Sound as a picture

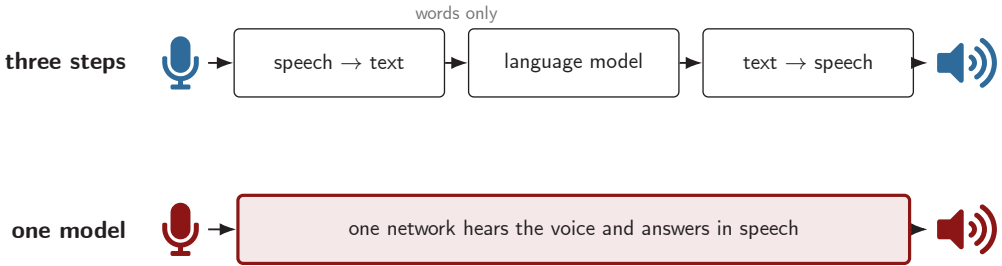


A synthetic voice (Kokoro), drawn on this laptop. Bright: loud at that pitch and moment.

- Sound: 16,000 numbers a second.
- Every 0.01 second: strength in 80 frequency bands.
- Cut into slices, one token each.
- Whisper: 680,000 hours of audio (78 years).

Source: Radford et al. (2023), "Robust Speech Recognition via Large-Scale Weak Supervision," pp. 1–3 (78 years: our arithmetic).

Speech in, speech out

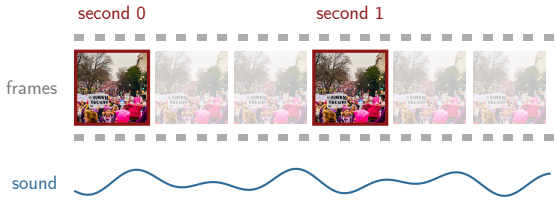


- Three steps: the middle model gets only the words.
- GPT-4o (2024): starts answering in about 0.3 seconds.
- 2026: models that listen while they speak.

Source: OpenAI, GPT-4o System Card (2024), p. 1; OpenAI, "Voice agents"; Google, Live API (checked October 9, 2026).

Watching a video

- A video: frames plus a soundtrack.
- Gemini keeps 1 frame a second.
- About 100 tokens a second.
- An hour-long meeting: about 360,000.



each second: 66 tokens for the frame
+ 32 for the sound ≈ 100

Illustration: real videos often have about 30 frames a second. Fast action between kept frames can be missed.

Source: Google, Gemini API docs, "Video understanding" (last updated October 9, 2026): 1 frame a second, 66 tokens a frame at the default (low) resolution, 32 a second for audio. 3,600 seconds × 100: our arithmetic.

Activity, part 2: is that what you said?

Same groups. One laptop per group. No login: it runs on the laptop.

musiclab.chromeexperiments.com/Spectrogram

1. Say "sssss," "ahhh," then whistle. Which one is a single thin line?

huggingface.co/spaces/Xenova/whisper-web

2. Each person reads one sentence. Copy what Whisper wrote.
3. Count the wrong words (rules on the sheet).

Read aloud

A. Senator Gillibrand and Representative Ocasio-Cortez spoke about the filibuster.

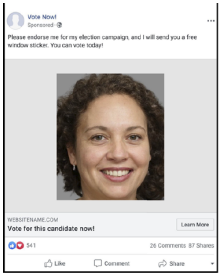
B. The Supreme Court decided Rucho versus Common Cause, a gerrymandering case.

C. In Maricopa County, election workers counted provisional ballots.

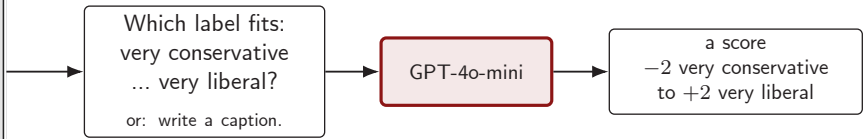
Chrome Music Lab (Google); Whisper Web (Xenova, Hugging Face), which runs the smallest Whisper model in the browser.

Models that see and hear politicians

Same ad, different face



one ad, 40 faces

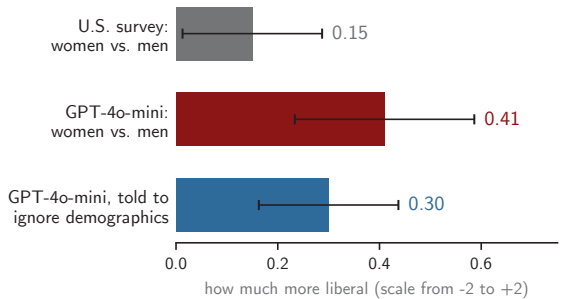


- One made-up Facebook ad; faces of Black and White women and men.
- 8,000 answers, each scored by GPT-4o.

Source: Jeon, Lee, Montgomery, and Lai (2026), *Political Analysis* 34, pp. 614–615. The ad: from the paper's Figure 1 (CC BY 4.0).

What the model read in the faces

- Women: 0.41 more liberal than men. U.S. survey: 0.15.
- Told to ignore race and gender: still 0.30.
- Race: about the survey's gap; smaller in GPT-4o.
- Gemini refused to rate faces.

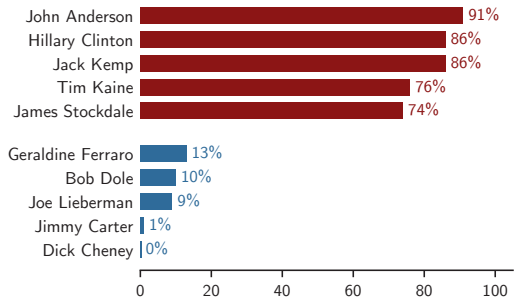


Lines: 95% confidence intervals.

Source: Jeon et al. (2026), Tables 1 and 2, pp. 617–619, and footnote 3, p. 614. Survey: Weidenbaum Center Survey (YouGov).

A speech model hears 60 years of debates

- All debates, 1960 to 2020: 67.5 hours.
- A speech model names the speaker: 95% (34 candidates).
- Emotion model's "angriest": Clinton second; Trump not in its top five.
- Hand coders, 2016 debate: Trump angry or threatening 83%, Clinton 35%.



The five candidates with the highest and the lowest share of clips labeled "angry."

Source: Mestre and Ryan (2026), *Political Analysis* 34, pp. 314, 321–324; Figure 4b redrawn, values read off the figure. Hand coding: Bucy et al. (2020), *New Media & Society* 22(4), p. 644.

Where these models fall short



Simple shapes

"Do two circles overlap?" and six other easy tasks: four top models (2024) right 58% of the time.

Rahmanzadehgervi et al. (2024)



Made-up words

Whisper (2023 tests): about 1% of transcripts had invented sentences; 38% of those contained harmful content.

Koenecke et al. (2024)



Fakes

Claude "cannot determine whether an image is AI-generated."

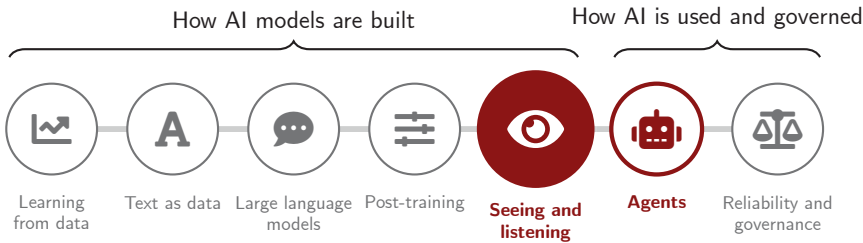
Anthropic, "Vision" (2026)

Discuss



1. A news site uses a model to describe candidates' photos. What should it check first?
2. Should a researcher code 100,000 protest photos with a model? What would you check, and how?
3. Back to today's question (think of the face study): when a model guesses a candidate's politics from a photo, what is it looking at?

Next week

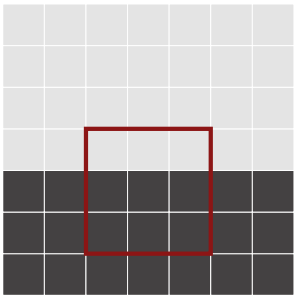


Next Wednesday, November 11

From chatbots to agents. Prep on Canvas: Anthropic, "Building effective agents" (first half), and a post by METR, a group that tests AI models.

Backup slides

Before transformers: filters that slide



a 7 × 7 image

+1	+1	+1
0	0	0
−1	−1	−1

the filter

- Multiply each pixel under the red square by the filter's number, then add.
- Slide one step and repeat. A big sum marks an edge.
- A CNN learns its filters from labeled photos (session 3: AlexNet, 2012).

Source: Illustration. Krizhevsky, Sutskever, and Hinton (2012), pp. 4, 6–7; the CS 3370 slides (Yuan Yuan, after Justin Johnson).

In the ViT study, more training photos helped

Pretraining photos	Size	ViT vs. a CNN
ImageNet contest set (1,000 classes)	1.3 million	a few points worse
Full ImageNet (21,000 classes)	14 million	close
JFT (Google's own)	303 million	better

- A CNN assumes that nearby pixels go together. A ViT has to learn that from data.
- Best result: 88.55% right on ImageNet's held-out photos, after pretraining on Google's in-house JFT data and more training on ImageNet.

Source: Dosovitskiy et al. (2021), pp. 1–2, 4, and 6–7, Table 2 and Figure 3.

Making pictures: removing noise



- To make a new picture, start from pure noise and remove it step by step. At each step, the text prompt decides which way to go.

Source: Illustration with the noise schedule of Ho, Jain, and Abbeel (2020), pp. 2, 5; Rombach et al. (2022), p. 1. Generated media in politics: session 7.

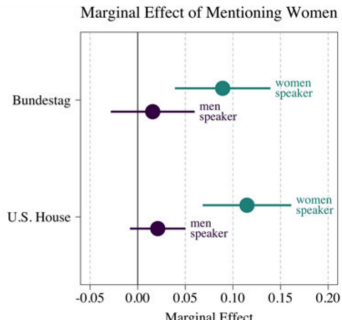
CLIP on all twelve photos

#	Photo	CLIP's top label	Score	Right?
1	March on Washington crowd, 1963	campaign rally	0.94	no
2	Women's March, signs, 2017	protest	0.91	yes
3	Suffragists' picket, White House, 1917	protest	0.52	yes
4	Sanders rally, Boston, 2020	campaign rally	0.86	yes
5	State of the Union, 2024	official portrait	0.85	no
6	A senator's official portrait	official portrait	0.99	yes
7	Early-voting line, Florida, 2008	polling place	0.97	yes
8	Inauguration crowd, 2009	campaign rally	0.54	unclear
9	Flag on a football field, 2010	sports game	0.94	yes
10	Music festival crowd, 2022	concert	0.99	yes
11	Candlelight vigil and protest, 2022	vigil	1.00	partly
12	Baseball fans, 1910–13	campaign rally	0.74	no

CLIP ViT-B/16, run on this laptop, October 10, 2026. Labels: "a photo of a" protest, campaign rally, concert, sports game, legislature, polling place, vigil; and "an official portrait." Photos: Wikimedia Commons; credits in the speaker notes.

Pitch in parliament

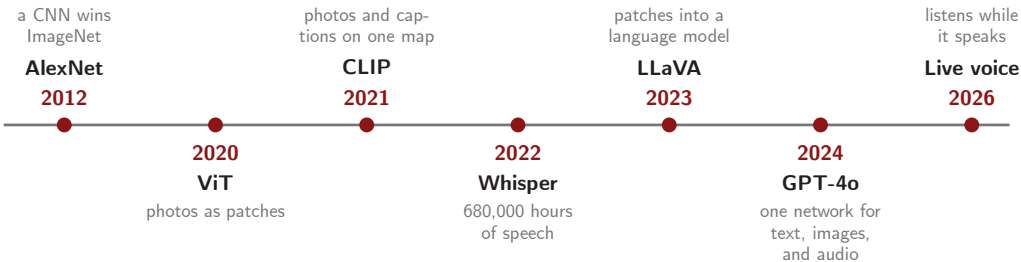
- German Bundestag, 2011 to 2020: 33,514 speeches with audio and text.
- Women mention women in 22% of their speeches; men in 11%.
- Speeches that use words about women: women's voices are higher by 0.089 standard deviations of their own pitch. Men's: 0.016, which could be zero.
- The U.S. House shows the same pattern.



Rise in pitch when a speech mentions women. Lines: 95% confidence intervals.

Source: Rittmann (2024), *British Journal of Political Science* 54(3), pp. 940–943, Figure 3; Dietrich, Hayes, and O'Brien (2019), *APSR* 113(4), p. 947.

From filters to models that listen



Source: Krizhevsky et al. (2012); Dosovitskiy et al. (2021, arXiv 2020); Radford et al. (2021); Radford et al. (2023, arXiv 2022); Liu et al. (2023); OpenAI (2024); OpenAI, "Voice agents" (2026).

Sources (1 of 2)

Anthropic (2026). "Vision." Claude API documentation, platform.claude.com, checked October 9.

Bucy, E. P., et al. (2020). "Performing populism: Trump's transgressive debate style and the dynamics of Twitter response." *New Media & Society* 22(4): 634–658.

Dietrich, B. J., Hayes, M., and O'Brien, D. Z. (2019). "Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech." *American Political Science Review* 113(4): 941–962.

Dosovitskiy, A., et al. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *ICLR* (arXiv:2010.11929).

Google (2026). "Media resolution" and "Live API" pages, Gemini API documentation, ai.google.dev, checked October 9.

Ho, J., Jain, A., and Abbeel, P. (2020). "Denoising Diffusion Probabilistic Models." *NeurIPS* 33.

Jeon, S., Lee, M. H. J., Montgomery, J. M., and Lai, C. K. (2026). "From Faces to Politics: Vision-Language Models (Sometimes) Link Visual Demographic Characteristics to Ideological Labels." *Political Analysis* 34: 613–621.

Koenecke, A., Choi, A. S. G., Mei, K. X., Schellmann, H., and Sloane, M. (2024). "Careless Whisper: Speech-to-Text Hallucination Harms." *ACM FAccT '24*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet Classification with Deep Convolutional Neural Networks." *NIPS* 25.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). "Visual Instruction Tuning." *NeurIPS* 36.

Mestre, R., and Ryan, M. (2026). "Potential and Pitfalls of Audio as Data for Political Research: Alignment, Features, and Classification Models." *Political Analysis* 34: 312–328.

Sources (2 of 2)

OpenAI (2024). "GPT-4o System Card," August 8 (arXiv:2410.21276).
OpenAI (2026). "Images and vision" and "Voice agents," API documentation, developers.openai.com, checked October 9. OpenAI Help Center, "Understanding and counting tokens."
Radford, A., et al. (2021). "Learning Transferable Visual Models From Natural Language Supervision." *ICML*, PMLR 139: 8748–8763.
Radford, A., et al. (2023). "Robust Speech Recognition via Large-Scale Weak Supervision." *ICML*, PMLR 202: 28492–28518.
Rahmanzadehgervi, P., Bolton, L., Taesiri, M. R., and Nguyen, A. T. (2024). "Vision language models are blind." *ACCV*: 18–34.
Rittmann, O. (2024). "Legislators' Emotional Engagement with Women's Issues: Gendered Patterns of Vocal Pitch in the German Bundestag." *British Journal of Political Science* 54(3): 937–945.
Rombach, R., et al. (2022). "High-Resolution Image Synthesis with Latent Diffusion Models." *CVPR*: 10684–10695.
Scholz, S., Weidmann, N. B., Steinert-Threlkeld, Z. C., Keremoğlu, E., and Goldlücke, B. (2025). "Improving Computer Vision Interpretability: Transparent Two-Level Classification for Complex Scenes." *Political Analysis* 33: 107–121.
Photos (Wikimedia Commons): Ted Eytan, Women's March, 2017, CC BY-SA 2.0, cropped and resized; the patch and noise images are adapted from it, CC BY-SA 2.0. Jp16103, Sanders rally, 2020, CC BY-SA 4.0. LatakiaHill, vigil, 2022, CC BY-SA 4.0. Public domain: National Archives (1963 march; 1917 picket), U.S. Congress, U.S. Air Force, U.S. Marine Corps, the White House, Library of Congress. Alex Johnson, early voting, CC BY-SA 2.0; PinkBeachPlanet, festival, CC BY-SA 4.0. Tools: Chrome Music Lab; Whisper Web (Xenova); a CLIP demo (merve/compare_clip_siglip); Arena (arena.ai), Battle mode and the Vision leaderboard (October 9, 2026 update).
Prep for Session 15: Schluntz, E., and Zhang, B. (2024). "Building effective agents." Anthropic, December 19; METR, metr.org/time-horizons.
Icons: Font Awesome Free 5 (CC BY 4.0). Video: Manim Community Edition; narration by Kokoro-82M, a synthetic voice.