AI's Visual and Musical Understanding of Chinese and American Cultures

Final Project, Yuehao Gao

MAT255, Fall 2024 | University of California, Santa Barbara

For this course's final project, I will scrutinize generative AI Models' understanding of the relationship between Chinese and American cultures. Specifically, the models will be asked to imagine pictures and music that demonstrate these two cultures during individual cultural events, personal scenarios, and the relationship between the two cultural groups during harmonious communicative occasions when the atmosphere is tense, also when both cultures are united to fight for a larger goal. The AI models will decide all the detailed scenarios, elements, and features of the generated pictures and music. The resulting pictures and music will be examined to see if they accurately capture the core of the cultures, demonstrate the atmosphere of the scene, or bring any unexpected elements or features.

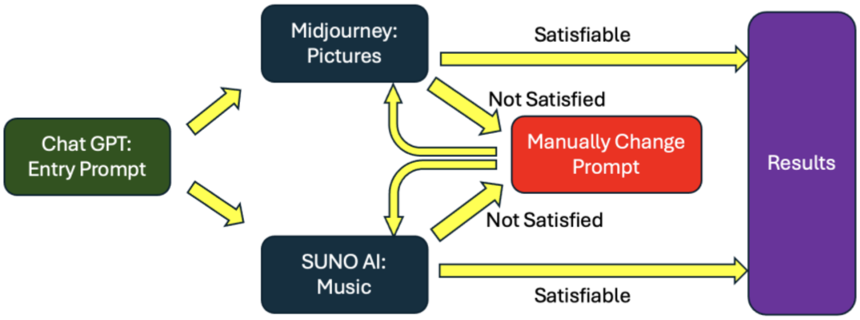
There will be three models utilized for this project:

(1). Chat GPT for imagining the prompts for generating the pictures and music

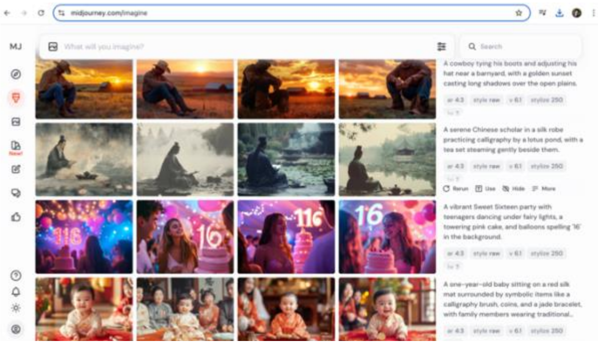
(2). Midjourney for the pictures

(3). SUNO AI for the music pieces

Specifically, the project starts with Chat GPT, so I asked it to generate the picture and music prompts for different scenarios separately. I have entered: *"Please imagine the prompts to AI pictures and music for the following scenes for Chinese and American cultures: during individual celebrations, individual cultural events, collaborative culture when having friendly relationships, cultures when having tense relationships, and collaborative cultures when collaborating to defend the intrude of aliens."* After Chat GPT exports the prompt for the pictures and music, I separately fed them into Midjourney and SUNO AI. If these generators' results are acceptable, they are then collected and analyzed. However, if the results are significantly diverged from the expected results, especially if due to lacking keywords in the prompt, then the new prompt will be entered to generate new results. The following is a workflow diagram:



This is to say, the Chat GPT is the major model being tested for imagining "what can be happening" in those 5 scenarios, respectively within Chinese cultures, American cultures, and merged cultures, while Midjourney and SUNO are tested more on their functions of illustrating the detailed visual or musical cultural elements. The following are the screenshots of the user interface of Midjourney and SUNO AI:

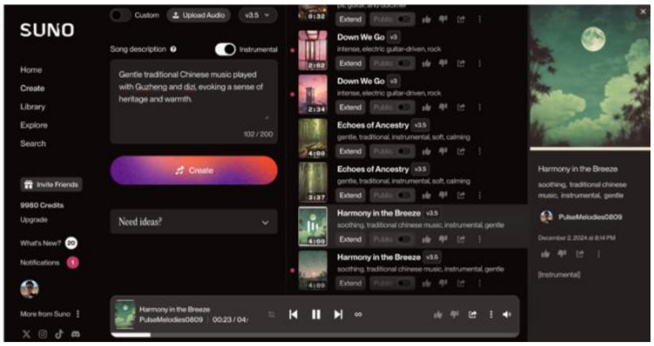


A cowboy tying his boots and adjusting his hat near a barnyard, with a golden sunset casting long shadows over the open plains.

A serene Chinese scholar in a silk robe practicing calligraphy by a lotus pond, with a tea set steaming gently beside them.

A vibrant Sweet Sixteen party with teenagers dancing under fairy lights, a towering pink cake, and balloons spelling '16' in the background.

A one-year-old baby sitting on a red silk mat surrounded by symbolic items like a calligraphy brush, coins, and a jade bracelet, with family members wearing traditional...



Gentle traditional Chinese music played with Guzheng and dizi, evoking a sense of heritage and warmth.

Since music files cannot be attached to this forum, they will be attached to the following website:

https://gaoyueh8.wixsite.com/home/ai-co-m-po

--- --- --- Scenario 1: Individual Celebrations --- --- ---

Prompt by Chat GPT:

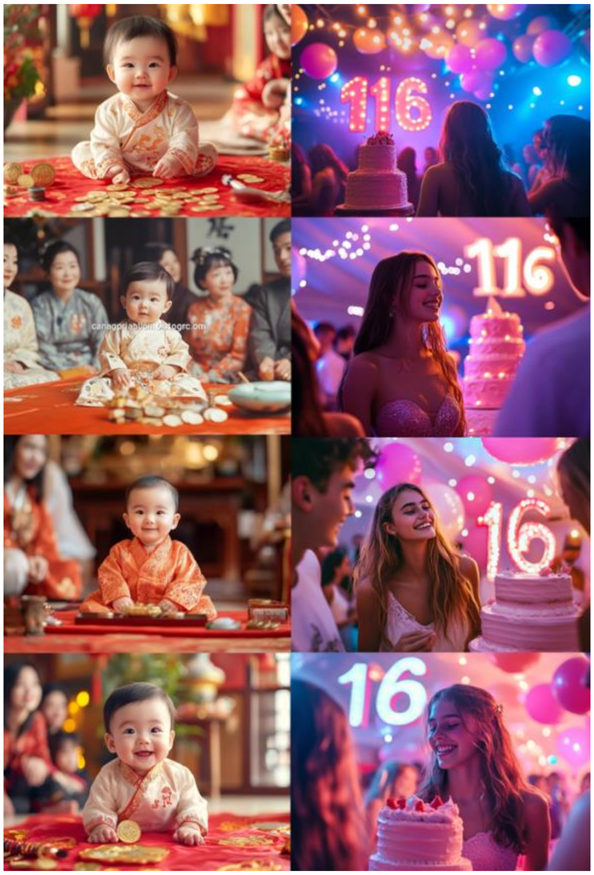
Picture for Chinese Context: A one-year-old baby sitting on a red silk mat surrounded by symbolic items like a calligraphy brush, coins, and a jade bracelet, with family members wearing traditional Chinese attire watching in anticipation.

Music for Chinese Context: Gentle traditional Guzheng music with soft flute melodies, evoking a sense of heritage and warmth.

Picture for American Context: A vibrant Sweet Sixteen party with teenagers dancing under fairy lights, a towering pink cake, and balloons spelling '16' in the background.

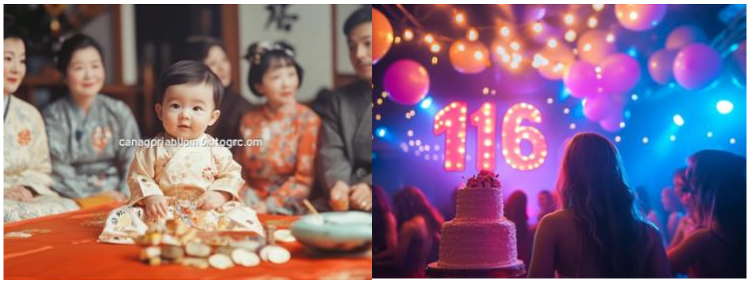
Music for American Context: Upbeat pop music with energetic rhythms and a prominent electronic beat, reflecting youthful celebration.

Generally, it is indispensable to say that birthdays are important individual celebrations for families and friends within both Chinese and American cultures. GPT has expressed the vibe and atmosphere in a clear, precise, and vivid way. The following are how Midjourney imagined the scenarios based on these texts:



The overall vibe accurately demonstrates the joyfulness when a baby is having his or her birthday, with family members cheerfully gathering around within Chinese cultures, or when a 16-year-old girl is gathering her friends in a much livelier atmosphere during her birthday celebration in American culture. Just as how most Chinese cultures overwhelm family over friends while American youth cultures may differ, the pictures accurately illustrate this point.

However, two pictures are significantly interesting: the second for the "Chinese cultures" has a line of letters that look like a "watermark". It is obvious that the samples that the Midjourney learned from, especially when searching for "Chinese babies", have a watermark in the picture. At the same time, the first picture for "American cultures" has the letters "116" instead of "16" on the cake, which further indicates that the picture model currently lacks the mechanisms that regulate the preciseness of word letters.



When SUNO AI generates the music for "Chinese individual celebrations", it is disappointing to find that both variations only consisted of light piano music: despite the C-major and the soft textures of the notes does create a warm family-like atmosphere, it is not played by traditional instruments like Guzheng as specified. Therefore, after I manually modified the prompt by adding a "Chinese" in the front, the new variations more precisely illustrated the scene with soft, warm, slow-tempo music that created even more feeling of "being born in an Eastern family".

While the two versions of "Neon Lights" start with soft and light piano melodies only, they gradually evolve to tense, dance-inducing, and exciting tones when evolving to the chorus parts. In particular, SUNO AI adds strong dance beats as well as electric MIDI notes to the music, boosting the partying vibe during the celebration of an American teenager's birthday.

--- --- --- Scenario 2: Individual Cultural Contexts --- --- ---

Prompt by Chat GPT:

Picture for Chinese Context: A serene Chinese scholar in a silk robe practicing calligraphy by a lotus pond, with a tea set steaming gently beside them.

Music for Chinese Context: Minimalist erhu and pipa music with flowing, meditative tones, emphasizing focus and tradition.

Picture for American Context: A cowboy tying his boots and adjusting his hat near a barnyard, with a golden sunset casting long shadows over the open plains.

Music for American Context: Country music with acoustic guitars and harmonicas, evoking a sense of rugged individualism and tradition.

For both cultures, Chat GPT decided to create a figure that demonstrates solitude and being alone. While it poured much more emphasis on its comprehension of Chinese culture being "poetic", with the characters dressed up in ancient costumes, it understood the American version of "solitude" of being lonely cowboys.

Both cultures involve the settings to be in nature: while Chinese involves mountains and bamboo forests, American involves large farmland and deserts.

This time, Midjourney precisely captures the scene for both cultures, as evidenced by a high congruence for all four variations of pictures for both cultures:

SUNO AI also performs better in this scenario. To illustrate the scene describing the Chinese poet being alone, it composed a piece titled "Quiet Heart", which consists of slow, quiet, and meditative melodies played by Guzheng (despite the MIDI system making it sound more like a Harp). The entire piece is in G major, with most notes being designed to fit into the Chinese pentatonic musical scale, which aligns with the acoustic features of Guzhengs in real life. Still, the sudden change in tone happening in the middle of the music is not achievable as Guzhengs are usually fixed to one tone within a single piece due to their sounding nature.

It becomes more interesting to investigate how SUNO composed country music with lyrics for the cowboy scenes. Generally, the use of major tones, drum sets, electric guitars, and bass guitars gives an uplifting and happy vibe. Although the lyric does show the theme of a "lonely road home", the harmonic progressions create a significant contrast. It may be the case that American country music generally tends to drag people out of depressive moods and boost hopes, which is a "hidden cultural context" that I did not expect before.

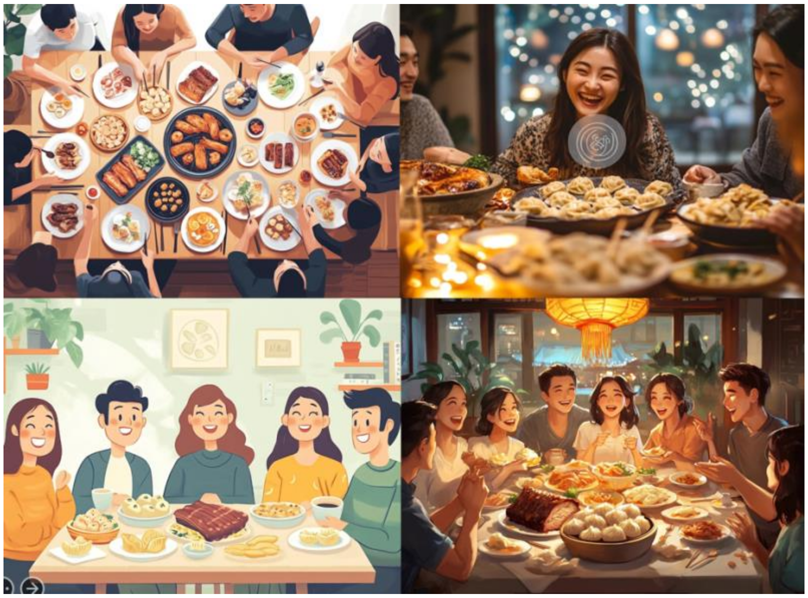
--- --- --- Scenario 3: Collaborative Cultures in Friendly Relationships --- --- ---

Prompt by Chat GPT:

Picture for Chinese-American Fusion Context: <u>A lively dinner table with dumplings, barbecue ribs, fortune cookies, and apple pie, surrounded by people laughing and chatting in a cozy dining room.</u>

Music for Chinese-American Fusion Context: <u>A fusion of traditional Chinese instruments and American jazz, creating a harmonious blend of melodies.</u>

Let's not worry about whether fortune cookies are traditional Chinese food or not. At least, it is obvious that the prompt lack of key descriptions of the ethics of people. Therefore, I upgraded the prompt for images by changing "people" to more specified "Chinese and American people".



Despite the command "--style raw" which asks the model to generate pictures in real-life styles, all the pictures that involve at least some sense of "Chinese people sitting around with Americans" are drawn in Cartoon styles. The only picture that is indeed in real-life style only consisted of Eastern figures, rather than a fusion of people from different cultural groups. This significantly indicates that Midjourney is lacking data for "joint celebration between Chinese and Americans" in its dataset. It could also be possible that relevant pictures do exist in Midjourney's database, but proper function labels like "joint", "fusion", "having fun together" and "cultural communication" are not given to those pictures, hindering the model from merging the data from two sets into one.

The first iteration for SUNO AI demonstrating the scene is also somewhat limited. The music begins with traditional pentatonic Guzheng notes only, then gradually proceeds into a pop-styled English song, but only with a very little sense of jazz, especially with a lack of semi-tones and eight-note bass notes. Hence, a newer prompt is given to SUNO AI: "<u>A jazz music played by traditional Chinese instruments.</u>" Nevertheless, it is amazing to see SUNO's combinational creativity this time. The music "Shanghai Serenade" features a jazz feeling in the background, while the major melodies are Guzheng and Dizi notes in pentatonic scales. This combination does create a happy and lively vibe for collaborative culture communication scenes.

--- --- --- Scenario 4: Cultures in Tense Relationships --- --- ---

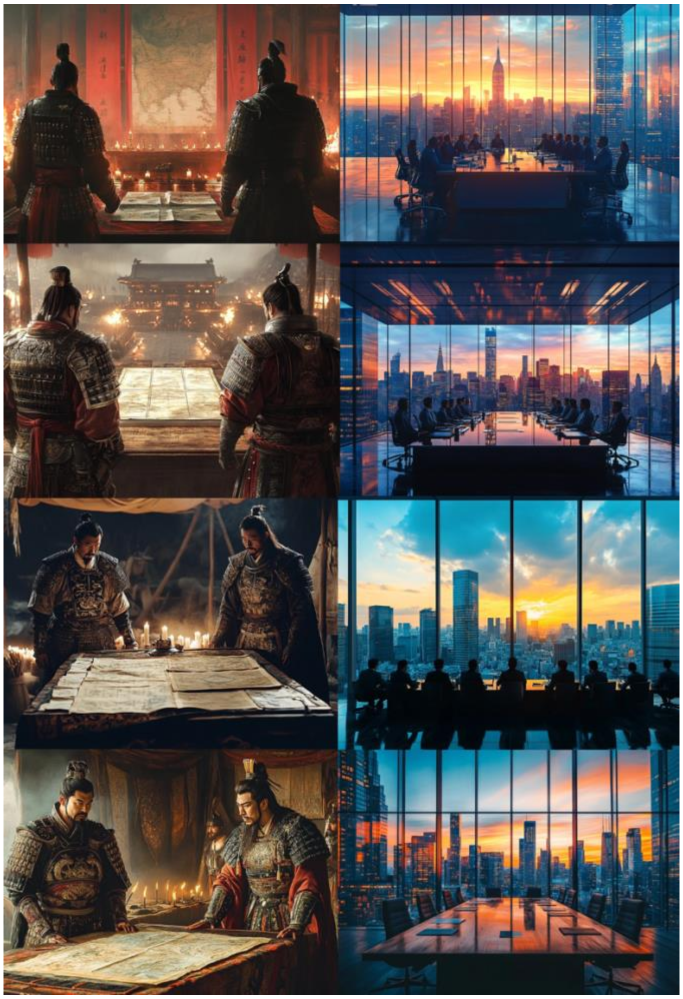
Prompt by Chat GPT:

Picture for Chinese Context: Two Chinese generals in ornate armor standing on opposite sides of a negotiation table in a candle-lit war tent, with maps and calligraphy scrolls spread out.

Music for Chinese Context: Tense guzheng and taiko drum patterns, creating a sense of conflict and anticipation.

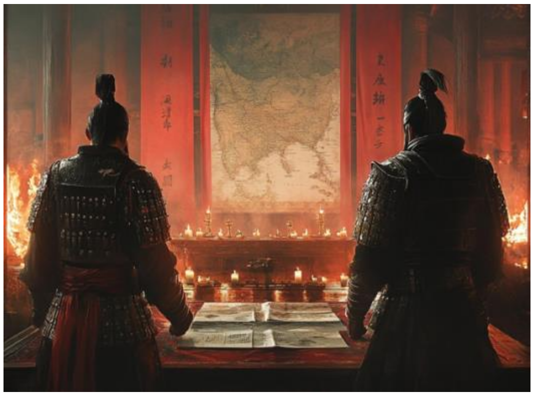
Picture for American Context: A modern corporate boardroom with sharp-suited executives locked in a heated debate, the skyline of a bustling city visible through the window.

Music for American Context: Low, pulsing electronic tones with sparse percussion, building a tense and serious atmosphere.

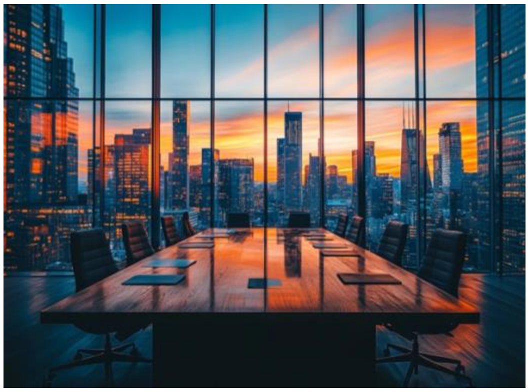


It is inevitable when two countries fall into tense relationships due to conflicts in benefits. As AI models may not generate scenarios for actual conflicts as they are not encouraged to advertise content that may hinder harmony, it is reasonable to imagine scenes where individual political or military leaders are having meetings to prepare for sudden unexpected events. The following are how Midjourney illustrated these scenes:

The hilarious thing is that all four iterations of "Chinese generals" are generals from hundreds of years ago. Despite that Google images display pictures of modern Chinese military generals, it is obvious that Midjourney is heavily "brainwashed" by television shows that illustrate ancient Chinese dynasties. While Chat GPT shall mean "a tent in a modern military field", Midjourney understood it as "a tent in an ancient battlefield", resulting in antique-themed pictures on the left.



For the American scenario, all pictures are set in a high-tech glass room of a tall building. As common sense tells me that military meetings are usually held in secret rooms with absolutely no windows, probably in the Pentagon, it also indicated that Midjourney has possibly gathered a lot of data samples from movies and television shows: these artistic works tend to illustrate American military and technology meetings in tall glass buildings, and "brainwashed" Midjourney as well. One exception is the fourth variation, where no human characters are presented in the entire canvas.



While SUNO AI successfully creates a tense atmosphere in music for both Chinese and American scenarios. The Chinese illustration creates a Guzheng melody that features a fast-tempo D-minor tone that shrinks the interval little by little, giving the sense that a storm or danger is slowly approaching and hiding somewhere around. The fast percussive instrument in the back also demonstrated this sense in general. However, SUNO AI did not successfully create a tense atmosphere describing the tension for the American cultures: the use of piano and electrical MIDI notes does illustrate the theme that the people discussing the military and politics are elites, but the melodies are gradually turning up, while the "i-VI-III-V" chordal progression features the sense of strength. However, this loop is more known to illustrate strengths for positive moods and hope and is not suitable for this tense scenario.

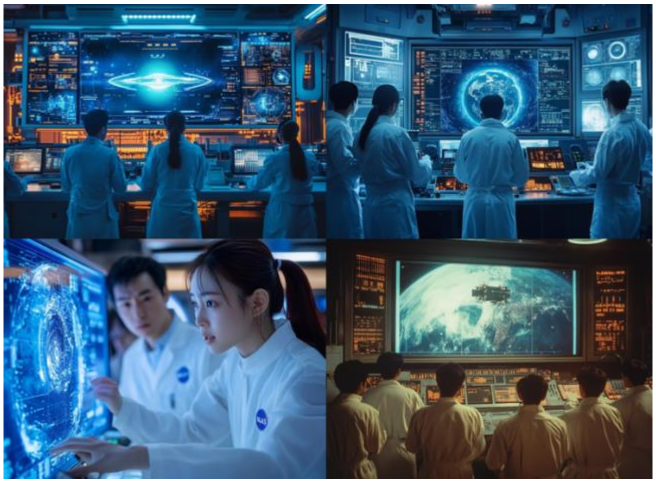
--- --- --- Scenario 5: Collaborative Cultures to Defend Against Alien Intrusion --- --- ---

Prompt by Chat GPT:

Picture Prompt: Chinese scientists in futuristic lab coats working with American NASA mission control scientists, gathered around a central screen showing an alien spacecraft approaching Earth.

Music Prompt: Majestic orchestral themes with brass, strings, and Chinese instruments, underscored by electronic synthesizers for a sci-fi edge.

It is highly reasonable for GPT or any human beings to think of the scenario of scientists from all countries gathering together to study how to defend against the intrude of aliens. These are how Midjourney imagined such a scene:



Similar to the problem in Scenario 3, the pictures lack "fusion" between Chinese and American characters. This could also be due to the lack of data or the lack of functioning labels that drive the model to link Chinese scientists and American scientists together in a collaborative scenario and work toward a collaborative goal.

SUNO AI faces a similar challenge when illustrating this scenario. Specifically, while it successfully incorporated electrical MIDI notes, which are arranged in fast descending or ascending F-minor triads, with a strong drum loop joining at the chorus part, which boosts the sense of high-tech and tension at the same time, it just completely ignored Chinese musical elements like Guzheng and pentatonic notes. While it is indeed hard to combine all the musical elements in one piece, SUNO has succeeded in Scenario 3. The reason why may be that key cultural words like "Chinese" should appear in the very front of the prompt, so that the NLP algorithm will increase its weight on the process of creating the music.

------- Conclusion --------

In general, all the AI models have performed well in imagining Chinese and American cultural scenarios. While Chat GPT performed well in understanding the basic cultures, it sometimes misses crucial words that highlight the significance of cultural elements, which shall lead to inaccuracies for Midjourney and SUNO. Midjourney is strong in illustrating what is given, but its imagination based on data from movies and TV shows, together with its lack of "combinational creativity," can limit its performance in cross-cultural content. Finally, while SUNO AI performs well in most scenarios, it is highly rigorous in the sequence of the keywords and is sensitive to the number of elements in its input prompt. Still, despite the minor flaws in these models, they are highly helpful for generating the pictures and music pieces for this project, as they have shown astounding progress in deep-learning algorithms. I am highly looking forward to seeing their future developments and how they shall boost positive intercultural understanding in the coming years.