Human Evaluation of I2MoE Interpretability

Background

I2MoE (Interpretable Multimodal Interaction-aware Mixture-of-Experts) learns and accounts for multi-modal interactions when making predictions. It does this by using different interaction experts that each learn a different type of interaction. For example, in the MM-IMDB dataset, I2MoE leveraged movie posters images and plot description text to predict movie genre. The **three main interaction types** are the following:

- **Uniqueness:** Unique information contained in each individual modality (e.g. image or text contain unique information useful for predicting genre)
- **Synergy:** Combination of multiple modalities can give rise to new information (e.g. when considered together, unique information from image and text combine to provide more insight into the genre)
- **Redundancy:** Shared information across the modalities (e.g. image and text contain similar information regarding genre)

I2MoE assigns weights to these interaction experts based on their contribution to the prediction.

Your Task

- You will be shown samples from the MM-IMDB dataset consisting of: (1) movie poster image, (2) plot description, and (3) the true genre of the movie.
- You will also be shown a bar plot of the weights assigned to each type of interaction (uniqueness of image, uniqueness of language, redundancy, synergy).
- Based on these four pieces of information, you will evaluate how well the model-assigned interaction weights make sense (e.g., if there is a great degree of overlapping information between the image and text pointing towards the genre, such as if the image and text contain same information for genre prediction, does the model assign more weight to redundancy?).
- Select on a scale from makes no sense at all to completely makes sense.

Illustrative examples are shown below in Figure 1 and Figure 2.

* Indicates required question

Figure 1. Example of uniqueness, synergy, and redundancy for a sample from MM-IMDB.

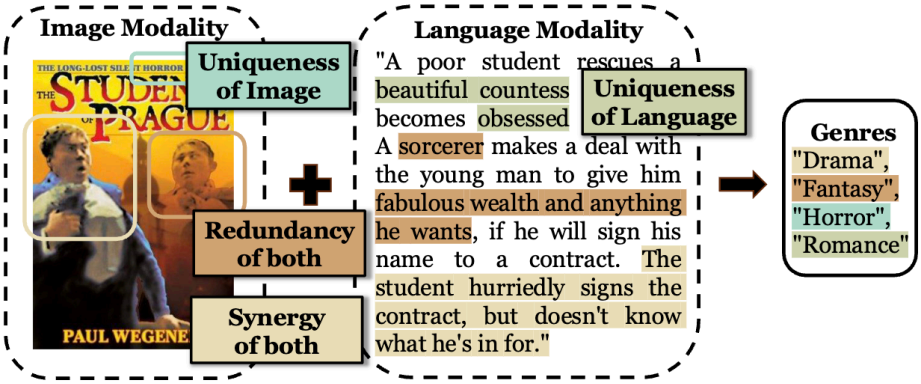


Figure 1. Example of uniqueness, synergy, and redundancy for a sample from MM-IMDB.

- The ground truth genres for this movie are Drama, Fantasy, Horror and Romance.
- **Image Uniqueness:** The word, 'HORROR' in the image is evidence of the Horror genre.
- **Language Uniqueness:** The words, 'beautiful', 'countess', and 'obsessed' in the text uniquely point towards the Romance genre.
- **Redundancy:** Both the image of the sorcerer (brown box) and the words 'sorcerer' and 'fabulous wealth and anything he wants' in the text point towards the Fantasy genre.
- **Synergy:** The image of the man with a surprised expression (tan box) and the text stating 'the student hurriedly signs the contract, but doesn't know what he's in for' are evidence unique to each modality that together provide evidence of the Drama genre.

Figure 2. Example of model output.

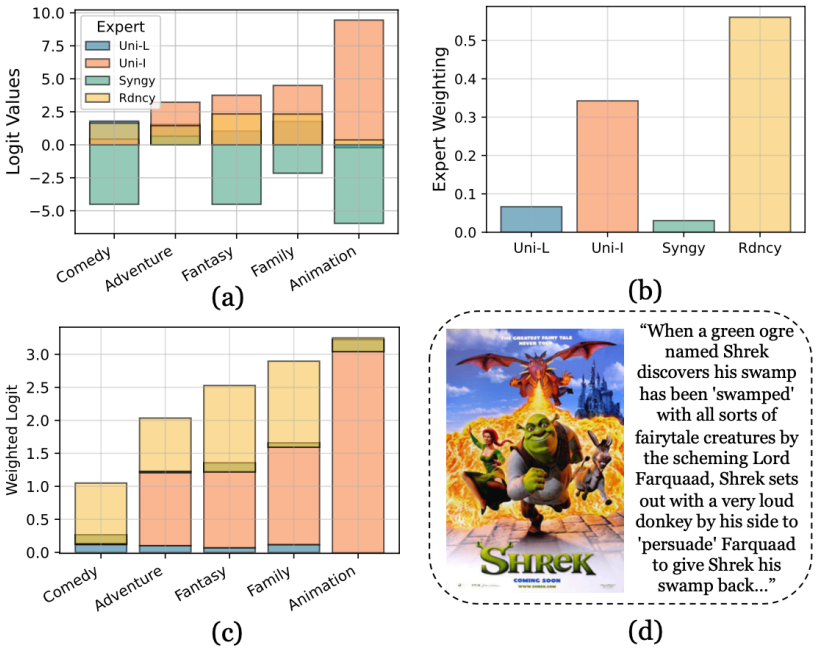


Figure 3. Qualitative example of local interpretation on the MM-IMDB dataset provided by `I²MoE-MulT`. Ground truth labels are `Comedy`, `Adventure`, `Fantasy`, `Family`, and `Animation`. (a) Logits output by different interaction experts. (b) Weighting assigned by the reweighting model. (c) Contribution of each interaction expert to the final weighted logit. (d) Raw image and language modalities used for prediction.

Figure 2. Example of model output.

- You will also be shown the graph of weights (panel (b) above) that the model assigned to each type of interaction. The y-axis is the model weight assigned to each type of interaction, and the x-axis is the type of interaction: **Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy**
- For the example above, the ground truth genres are Comedy, Adventure, Fantasy, Family, and Animation.
- The expert weighting in panel (b) suggests that the unique information in the imaging modality and the redundancy between the information provided by the image and text were most important for determining the predicted genres.
- This makes sense, as the text largely describes what is depicted in the image, so it is expected that there is a high degree of redundancy. Furthermore, the image contains some unique information, such as the cartoon characters, bright colors, and silly facial expressions which uniquely provide evidence of the Animation, Comedy, and Family genre (for example, the text description does not indicate that the movie is animated).

First Name *

Your answer

Last Name *

Your answer

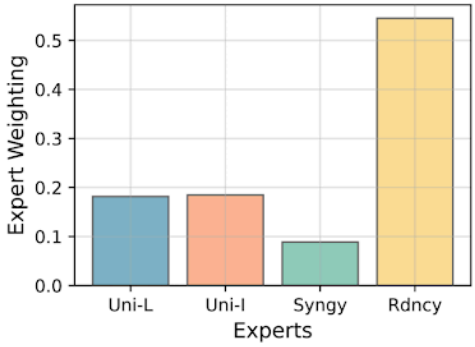
Email Address *

Your answer

Image 1/20: *

Ground Truth Genres: Drama, War

Text: "During 1st WW, two French officers are captured. Captain De Boeldieu is an aristocrat while Lieutenant Marechal was a mechanic in civilian life. They meet other prisoners from various backgrounds, as Rosenthal, son of wealthy Jewish bankers. They are separated from Rosenthal before managing to escape. A few months later, they meet again in a fortress commanded by the aristocrat Van Rauffenstein. De Boeldieu strikes up a friendship with him but Marechal and Rosenthal still want to escape..."



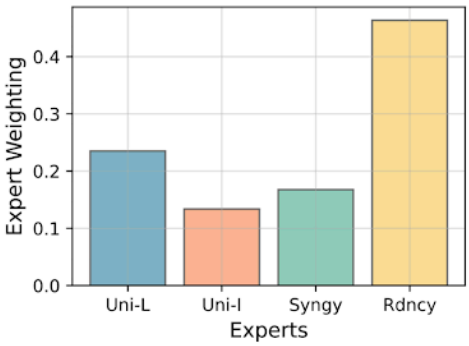
(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Image 2/20: *

Ground Truth Genres: Action, Adventure, Sci-Fi

Text: "Through a series of coincidences, Carrie, Dan and Dr. Hartmann all fall through a teleporter device Hartmann has invented. Transported to a what appears to be a prehistoric world in a parallel universe and unable to find the Doctor, Dan and Carrie must figure out a way to get back home. Before they can do that, however, they must deal with tribes of savage cavemen, as well as brutal warlord named Kleel who has taken a liking to Carrie and seems to be unusually well-supplied with Earth technology."



(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Image 3/20: *

Ground Truth Genres: Comedy, Adventure, Fantasy, Family, Animation

Text: "The Care Bears live in a country high in the clouds, where they have a lot of fun together. But they also do care for the human children on Earth, who they watch through huge telescopes from the sky, and come to help whenever there is need. Nikolas, a magician's apprentice, is in danger of getting under the influence of a bad spirit, which resides in an ancient spell book. The siblings Kim and Jason don't trust anyone anymore after being disappointed once too often. The Care Bears take them into their wonderland where they experience exciting and dangerous adventures together and quickly become good friends."

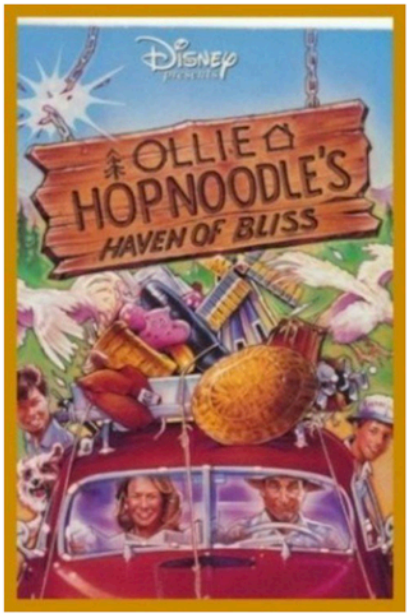
(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

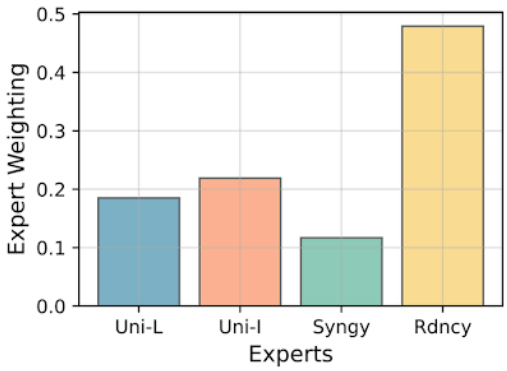
- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Image 4/20: *

Ground Truth Genres: Comedy, Family

Text: "It's summertime in Hohman, Indiana, and 14-year-old Ralph Parker can't wait to get his first job. His friends Schwartz and Flick are less enthusiastic, and the job turns into a nightmare presided over by the story's author, Jean Shepherd, in hilariously unconvincing movie makeup. Ralph's dad hardly can wait for the family's upcoming fishing vacation in Michigan at the movie title's resort but will have to drive their old car through a sea of troubles to get there. Fuzzhead the dog runs away and joins a wealthy family, to the consternation of Mom, who patiently handles über-whiner little-brother Randy and buys a whirligig as consolation."





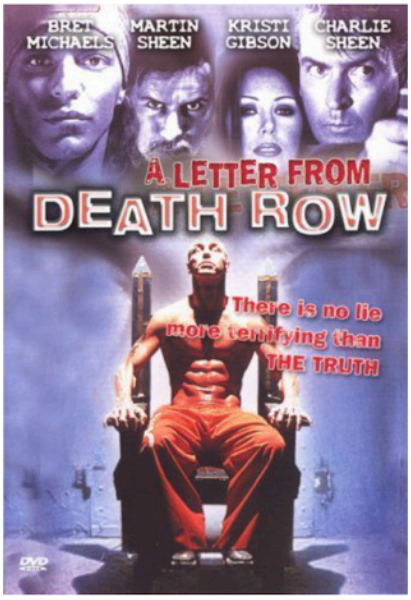
(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

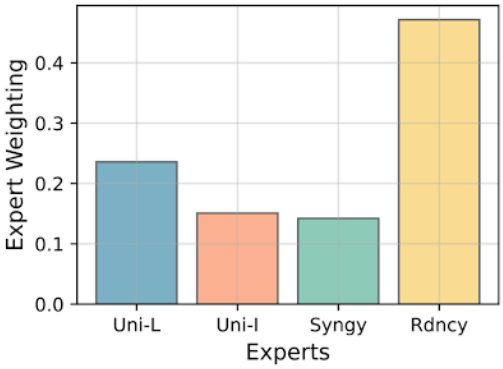
- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Image 5/20: *

Ground Truth Genres: Drama, Thriller, Crime, Mystery

Text: "A psychological thriller (...or is it?) that takes you through the mind of convicted killer, Michael Raine, and his experiences on death row... or does it? Was he guilty of killing his girlfriend or was he a victim of a conspiracy to frame him for a crime he didn't commit? As the story unfolds...or does it? Jessica Foster, an assistant to the Governor of Tennessee begins to interview Raine while on death row, claiming that she's writing a book about the inmates...Or is she? Through various circumstances, Raine puts two and two together and builds a case that he believes can prove his innocence...or does he? Ms. Foster is the only one on the 'outside" who can give Raine a voice, but is she working for those who framed him? As time draws near to the date of his execution, in his most desperate hour Raine finds the missing pieces to the puzzle to prove his innocence, but is it too late...? Was this story told from Raine's point of view or from the book writers or from yours, the viewer - you decide. All I know is - Damn I want that defense attorney when I'm on trial! Part of the Death Row Universe. Nominated for the little_cc 'Movie of the Decade' & loved among all 5 Poison fans."





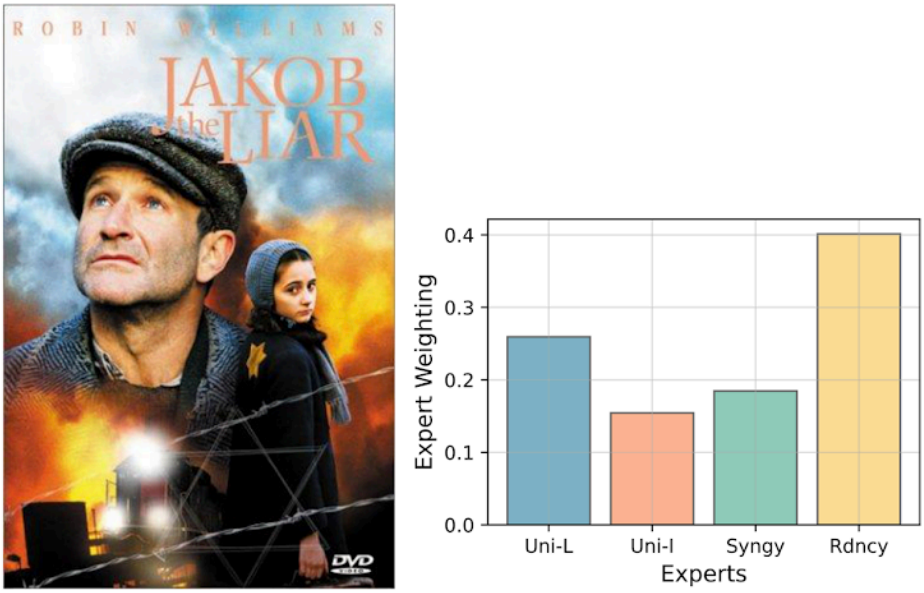
(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Image 6/20: *

Ground Truth Genres: Drama, War

Text: "In 1944 Poland, a Jewish shop keeper named Jakob is summoned to ghetto headquarters after being caught out near curfew. While waiting for the German Kommondant, Jakob overhears a German radio broadcast about Russian troop movements. Returned to the ghetto, the shopkeeper shares his information with a friend and then rumors fly that there is a secret radio within the ghetto. Jakob uses the chance to spread hope throughout the ghetto by continuing to tell favorable tales of information from "his secret radio." Jakob, however, has a real secret in that he is hiding a young Jewish girl who escaped from a camp transport train. A rather uplifting and slightly humorous film about World War II Jewish Ghetto life."



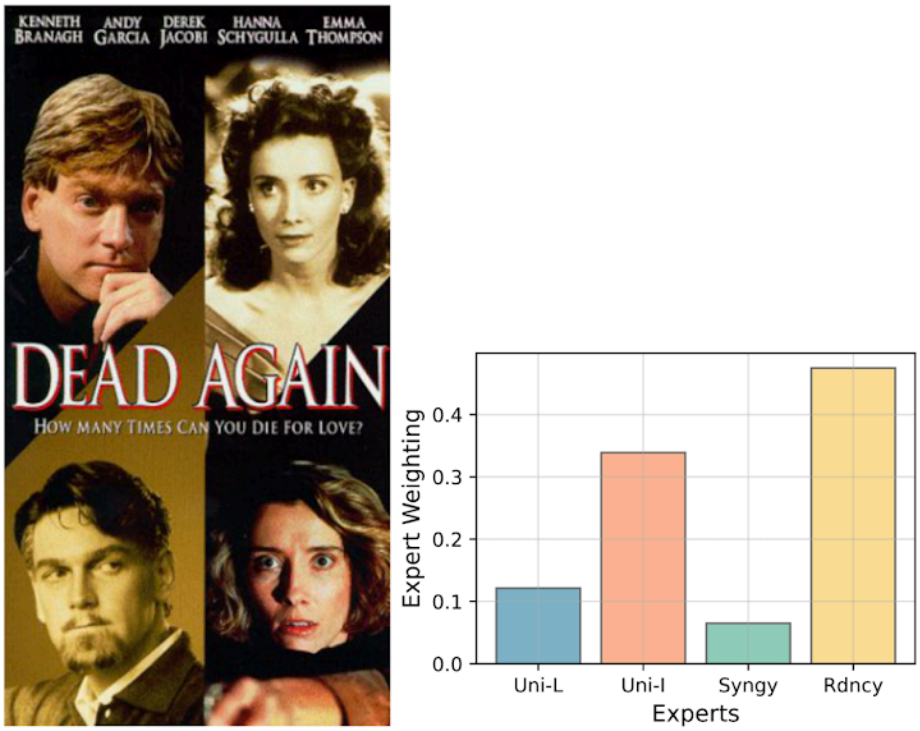
(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Image 7/20: *

Ground Truth Genres: Drama, Thriller, Crime, Mystery

Text: "Mike Church is a Los Angeles private detective who specializes in finding missing persons. He takes on the case of a mystery woman whom he calls Grace. She is suffering from amnesia and has no memories of her own. She keeps having nightmares involving the murder of a pianist, Margaret, by her husband Roman Strauss in the late 1940s. In an attempt to solve the mystery about these nightmares, Church seeks the help of Madson who is an antiques dealer with the gift of hypnosis. The hypnosis sessions will soon begin to reveal some surprises."



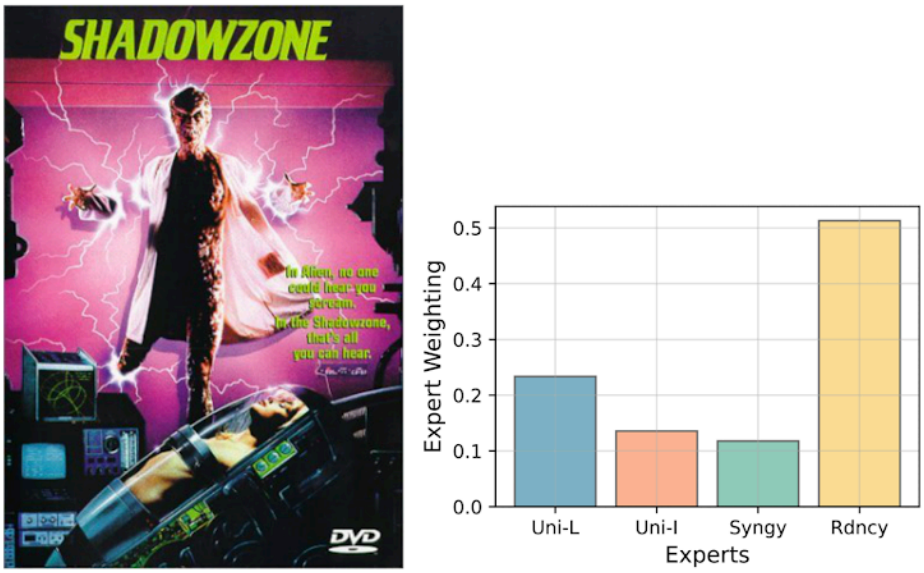
(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Image 8/20: *

Ground Truth Genres: Horror, Sci-Fi

Text: "After someone is killed in the subterranean project called "Shadowzone," a NASA captain is called in to investigate. In the project, sleeping subjects are induced into a deep EDS state whereby they become portals to a parallel universe. Unfortunately this causes adverse reactions in the subject, and something gets through the portal, the consequence of which is an attrition problem."



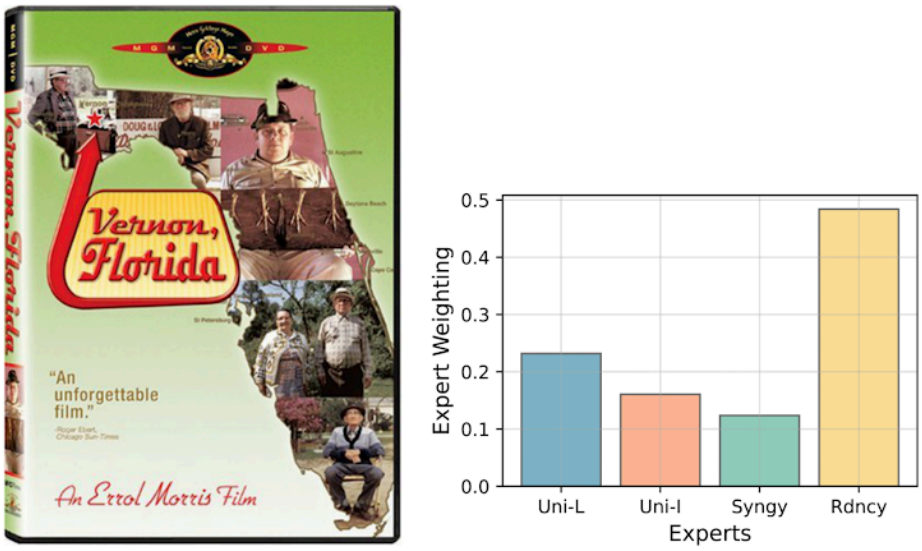
(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Image 9/20: *

Ground Truth Genres: Comedy, Documentary

Text: "Early Errol Morris documentary intersplices random chatter he captured on film of the genuinely eccentric residents of Vernon, Florida. A few examples? The preacher giving a sermon on the definition of the word "Therefore," and the obsessive turkey hunter who speaks reverentially of the "gobblers" he likes to track down and kill."



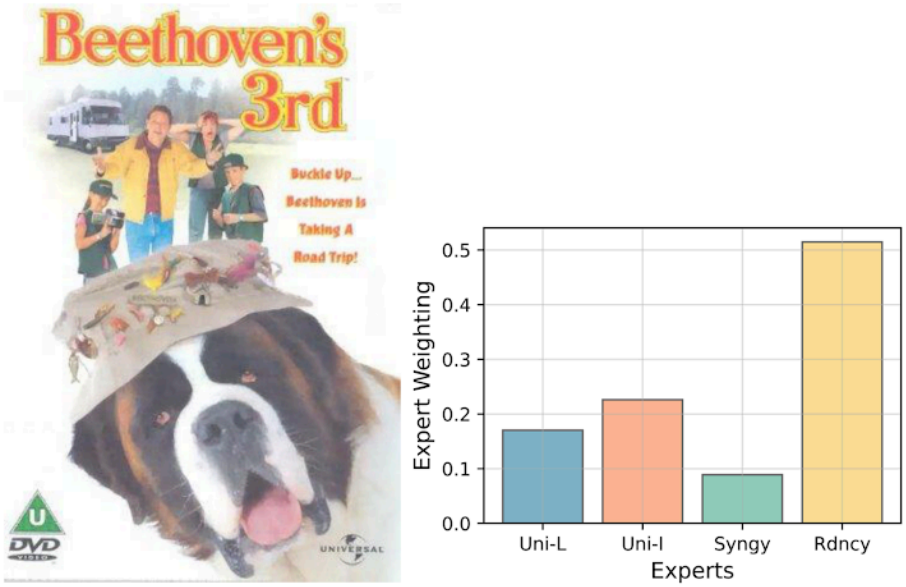
(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Image 10/20: *

Ground Truth Genres: Comedy, Family

Text: "The Newton family from the original Beethoven movies are on vacation in Europe but do plan to join a Newton family reunion and to make sure one of their family members definitely makes it, they ship him to travel to the reunion with George Newton's brother Richard. Guess which family member it was? That's right, Beethoven! The giant mutt follows Richard Newton and his family of a nagging wife and two bratty kids as they hit the road to California in a huge, shiny - and expensive RV, equipped with a DVD player. Following them are two bumbling crooks who have hidden some secret codes on a DVD that they figure no one in the world will buy, but someone does: Richard. So now they've got a DVD holding top secret information and the crooks must get it back..."



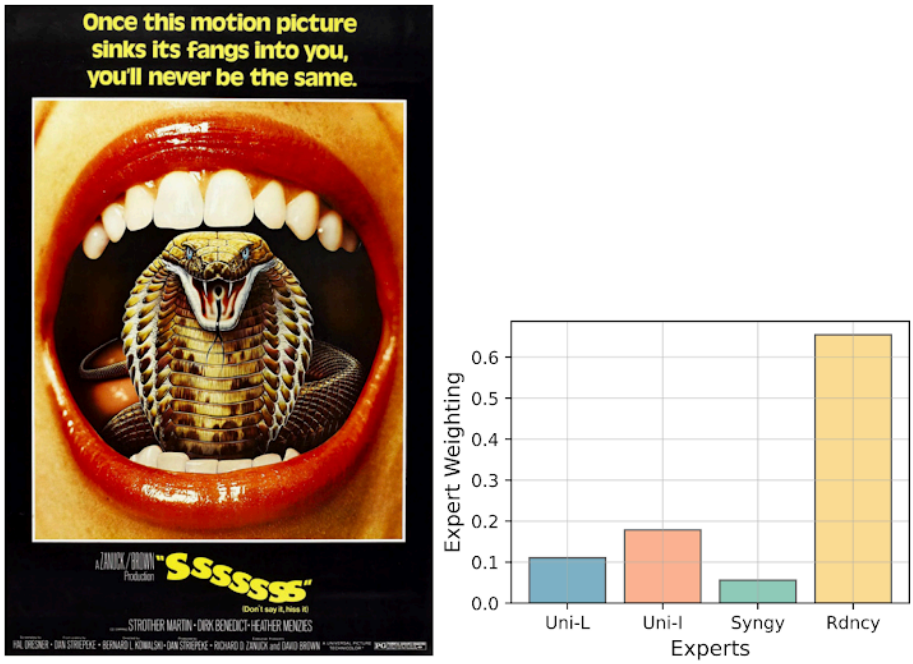
(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Image 11/20:

Ground Truth Genres: Horror, Sci-Fi

Text: "David, a college student, is looking for a job. He is hired by Dr. Stoner as a lab assistant for his research and experiments on snakes. David also begins to fall for Stoner's young daughter, Kristina. However, the good doctor has secretly brewed up a serum that can transform any man into a King Cobra snake-and he plans to use it on David."



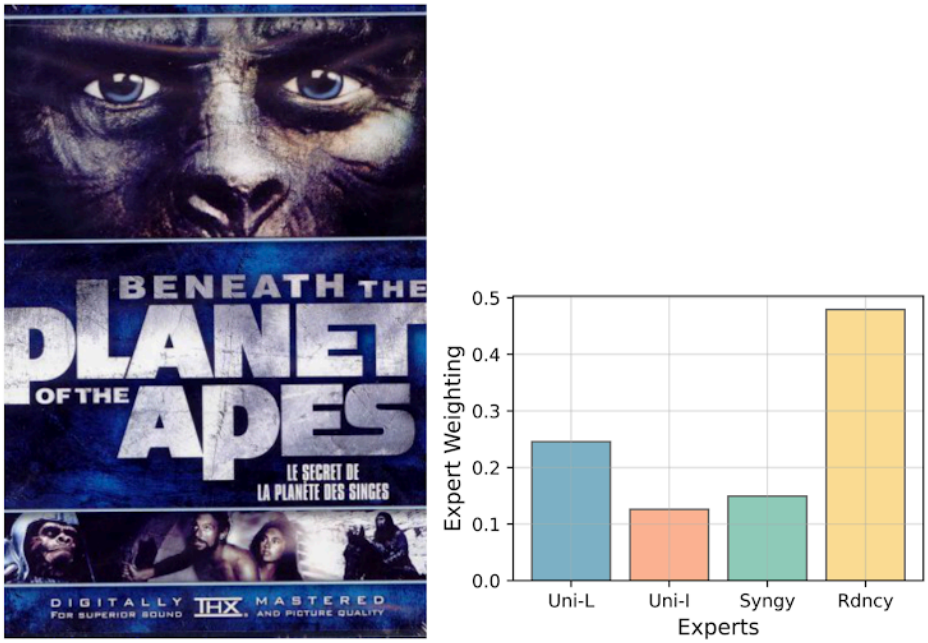
(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Image 12/20: *

Ground Truth Genres: Action, Adventure, Sci-Fi

Text: "Brent is an American astronaut, part of a team sent to locate missing fellow American astronaut, George Taylor. Following Taylor's known flight trajectory, the search and rescue team crash lands on an unknown planet much like Earth in the year 3955, with Brent being the only survivor of the team. What Brent initially does not know, much like Taylor didn't initially know when he landed here before Brent, is that he has landed back on Earth in the future, in the vicinity of what was New York City. Brent finds evidence that Taylor has been on the planet. In Brent's search for Taylor, he finds that the planet is run by a barbaric race of English speaking apes, whose mission is in part to annihilate the human race. Brent eventually locates some of those humans, who communicate telepathically and who live underground to prevent detection by the apes. These humans, who are in their own way as barbaric as the apes, want in turn to protect their species. Brent has to figure out a way to save himself under the circumstances, which may be more difficult to accomplish in the battle between the dominant species on this planet."



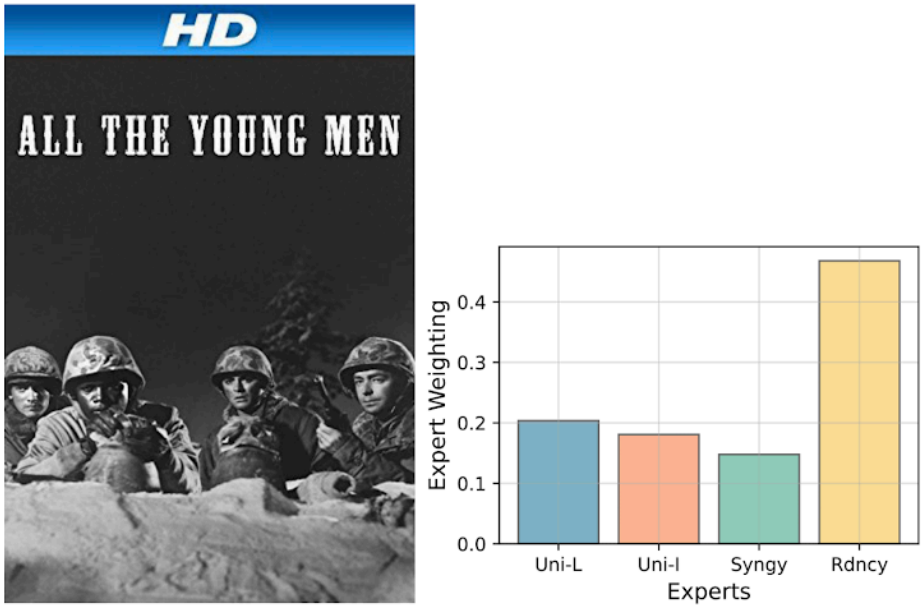
(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Image 13/20: *

Ground Truth Genres: Drama, War

Text: "During the Korean War, the lieutenant in charge of a Marine rifle platoon is killed in battle. Before he dies, he places the platoon's sergeant, who's black, in charge. The sergeant figures on having trouble with two men in his platoon: a private who has much more combat experience than he does, and a racist Southerner who doesn't like blacks in the first place and has no intention of taking orders from one."



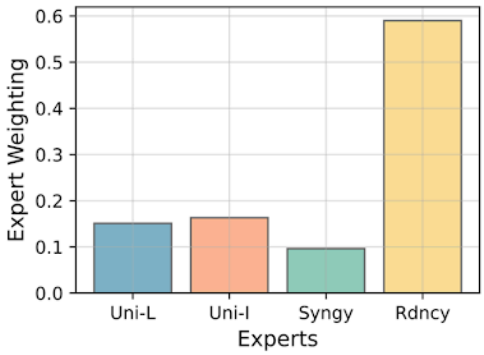
(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Image 14/20: *

Ground Truth Genres: Action, Adventure, Sci-Fi

Text: "The underground kingdom of Seatopia sends out Megalon, a giant beetle, and Gigan to destroy the above ground dwellers. In an attempt to stop them, an independently thinking robot brings Godzilla into the fight."



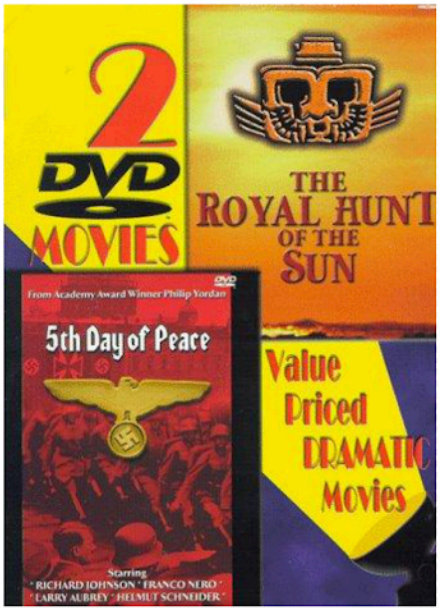
(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

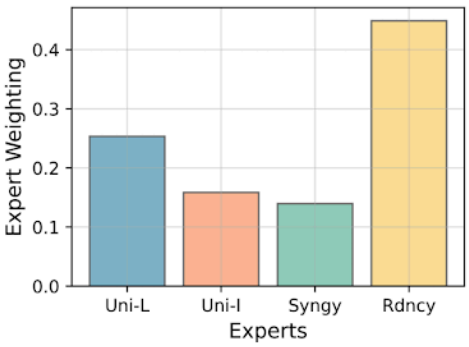
- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Image 15/20: *

Ground Truth Genres: Drama, War

Text: "At the end of WW II, German deserters are tried for desertion by fellow POWs inside a prisoner of war camp for Nazis."





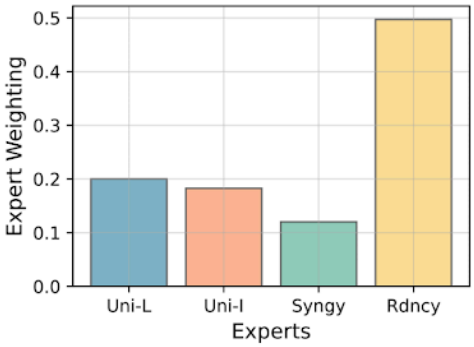
(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Image 16/20: *

Ground Truth Genres: Action, Adventure, Sci-Fi

Text: "It is the 23rd century. Admiral James T. Kirk is an instructor at Starfleet Academy and feeling old; the prospect of attending his ship, the USS Enterprise--now a training ship--on a two-week cadet cruise does not make him feel any younger. But the training cruise becomes a deadly serious mission when his nemesis Khan Noonien Singh--infamous conqueror from late 20th century Earth--appears after years of exile. Khan later revealed that the planet Ceti Alpha VI exploded, and shifted the orbit of the fifth planet as a Mars-like haven. He begins capturing Project Genesis, a top secret device holding the power of creation itself, and schemes the utter destruction of Kirk."



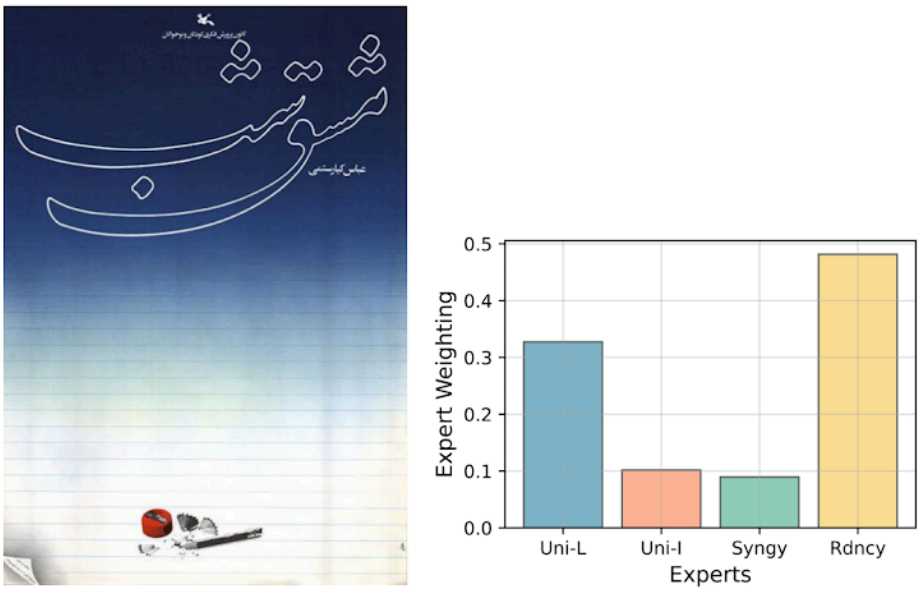
(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Image 17/20: *

Ground Truth Genres: Documentary

Text: "In this documentary, Kiarostami asks a number of students about their school homework. The answers of some children shows the darker side of this method of education."



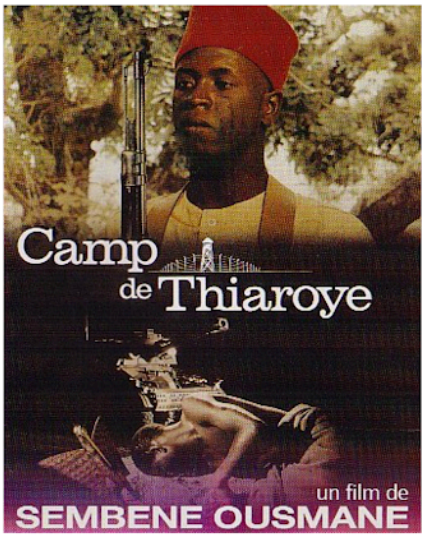
(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

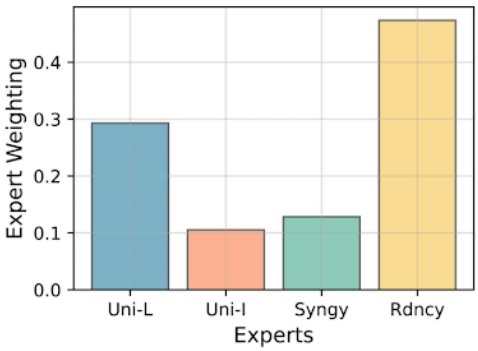
- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Image 18/20: *

Ground Truth Genres: Drama, War

Text: "In this semi-autobiographical film, black soldiers help to defend France, but are detained in prison camp before being repatriated home."





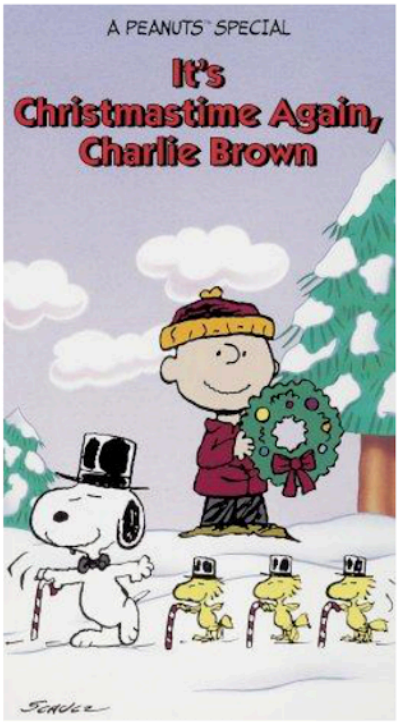
(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

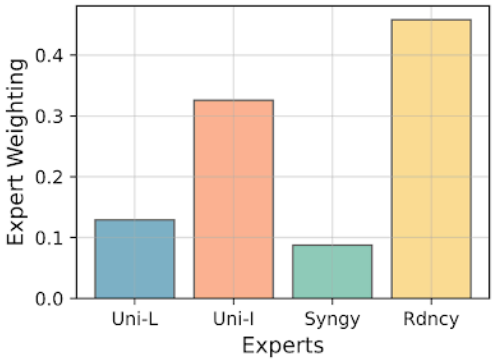
- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Image 19/20: *

Ground Truth Genres: Comedy, Family, Animation

Text: "As the holiday season rolls around and all the Peanuts gang are getting ready for it. Whether it be Charlie Brown struggling to raise money for his girlfriend or Sally and Peppermint Patty struggling to rehearse and memorize their one word lines for the Christmas pageant, these kids try to keep with the Christmas spirit while Snoopy has his mischief to do."





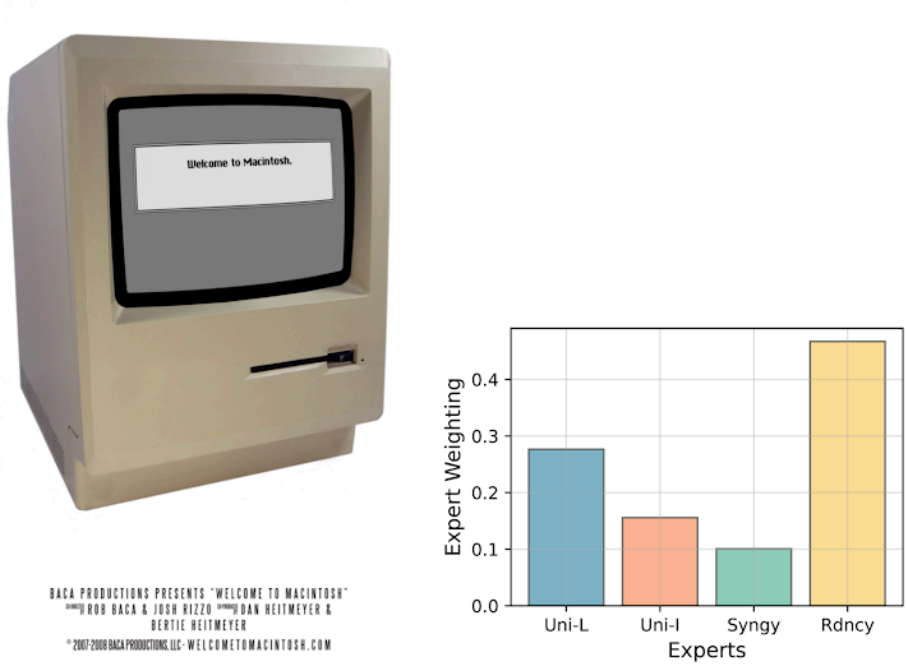
(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Image 20/20: *

Ground Truth Genres: Documentary

Text: ""Welcome to Macintosh" is a documentary that mixes history, criticism and an unapologetic revelry of all things Apple. Whether a long time Mac fanatic or new to computers, Welcome to Macintosh explores the many ways Apple Computer (now Apple, Inc.) has changed the world, from the early days of the Apple-I to the latest the company has to offer."



(left) Image modality; (right) Interaction Weights: y-axis is the model weight assigned to each type of interaction; x-axis is the type of interaction: Uni-L = unique information contained in the text (language) modality; Uni-I = unique information contained in the image modality; Syngy = synergy, and Rdncy = redundancy

- ◯ Completely makes sense
- ◯ Mostly makes sense
- ◯ Neutral
- ◯ Makes little sense
- ◯ Makes no sense at all

Submit Clear form