Supplementary Material of "Structured Policy Optimization: Enhance Large Vision-Language Model via Self-Referenced Dialogue"

Guohao Sun[1]*, **Can Qin**[2], **Yihao Feng**[2], **Zeyuan Chen**[2], **Ran Xu**[2], **Sohail Dianat**[1], **Majid Rabbani**[1], **Raghuveer Rao**[3], **and Zhiqiang Tao**[1]*

[1]Rochester Institute of Technology
[2]Salesforce AI Research [3]DEVCOM Army Research Laboratory

1. Overview

We provide the detailed derivation of SPO's objective function, more results and discussions, and qualitative examples of the proposed SPO.

1. Mathematical derivation of the proposed SPO. (Section 2)
2. More experimental settings (Section 3.1), model discussion (Section 3.2) and hyper-parameter analysis (Section 3.3) of our SFT-7B trained with SPO.
3. Qualitative results of real-world visual chat (Section 3.4).

2. Derivation

2.1. Sub-tree Reward Formulation

The sub-tree starting from *question* leads to two answers, which can be used to compute the preference probability. Intuitively, it follows the original DPO structure. Therefore, we can formulate the sub-tree reward $r(x, \mathcal{Y})$ as reward of *instruct* weighted by the probability of preference answer:

$$r(x, \mathcal{Y}) = r(x, q) * \sigma(r([x, q], y_w) - r([x, q], y_l)). \quad (1)$$

The preference probability $p(\mathcal{Y}_w \succ \mathcal{Y}_l \mid x)$ between the preferred sub-tree $\mathcal{Y}_w$ and the dispreferred sub-tree $\mathcal{Y}_l$ can be represented as

$$\begin{aligned} &p(\mathcal{Y}_w \succ \mathcal{Y}_l \mid x) \\ &= \sigma\big[r(x, q_w) * \sigma(r([x, q_w], y_w^1) - r([x, q_w], y_l^1)) \\ &- r(x, q_l) * \sigma(r([x, q_l], y_w^2) - r([x, q_l], y_l^2))\big], \end{aligned}$$

where $r(x, q) = \beta \log \frac{\pi_\theta(q|x)}{\pi_{\text{ref}}(q|x)} + \beta \log Z(x)$, $y_w^1, y_l^1 \sim \pi_{\text{ref}}(\cdot|x, q_w))]$ and $y_w^2, y_l^2 \sim \pi_{\text{ref}}(\cdot|x, q_l))]$. This objective is computationally expensive since $Z(x)$ is difficult to utilize. Therefore, we leverage trajectories to compute such structured reward in the following derivation.

*Corresponding authors: Guohao Sun (gs4288@rit.edu) and Zhiqiang Tao (zhiqiang.tao@rit.edu)

2.2. Deriving the SPO Objective Function Under the Bradley-Terry Model

The total reward of a trajectory is defined as $r(\tau) = \sum_{t=1}^{T} r(s_t, a_t)$ [2]. Following this reward formulation, we define our trajectory reward as

$$r(\tau) = \lambda \cdot r(x, q) + \delta \cdot r([x, q], y), \quad (2)$$

where λ and δ are hyper-parameters used to control the importance of question q and answer y preference.

Given the definition of trajectory reward, the loss function i.e., $\mathcal{L}(\tau_w, \tau_l)$, of classifying between preferred τ_w and dispreferred τ_l can be defined as

$$\begin{aligned} &-\mathbb{E}_{(\tau_w,\tau_l)\sim\mathcal{D}}\left[\log \sigma\left(r(\tau_w) - r(\tau_l)\right)\right] \\ =&\mathbb{E}_{(\tau_w,\tau_l)\sim\mathcal{D}}\Bigg[-\log\sigma\Bigg(\beta \log \frac{\pi_\theta(q_1|x)}{\pi_{\text{ref}}(q_1|x)} + \beta \log \frac{\pi_\theta(y_1|[x, q_1])}{\pi_{\text{ref}}(y_1|[x, q_1])} \\ &\qquad - \beta \log \frac{\pi_\theta(q_2|x)}{\pi_{\text{ref}}(q_2|x)} - \beta \log \frac{\pi_\theta(y_2|[x, q_2])}{\pi_{\text{ref}}(y_2|[x, q_2])}\Bigg)\Bigg] \\ =&\mathbb{E}_{(\tau_w,\tau_l)\sim\mathcal{D}}\Bigg[-\log\sigma\Bigg(\Bigg(\beta \log \frac{\pi_\theta(q_1|x)}{\pi_{\text{ref}}(q_1|x)} - \beta \log \frac{\pi_\theta(q_2|x)}{\pi_{\text{ref}}(q_2|x)}\Bigg) \\ &\quad + \Bigg(\beta \log \frac{\pi_\theta(y_1|[x, q_1])}{\pi_{\text{ref}}(y_1|[x, q_1])} - \beta \log \frac{\pi_\theta(y_2|[x, q_2])}{\pi_{\text{ref}}(y_2|[x, q_2])}\Bigg)\Bigg)\Bigg] \\ \le&\mathbb{E}_{(\tau_w,\tau_l)\sim\mathcal{D}}\Bigg[-\log\sigma\Bigg(\beta \log \frac{\pi_\theta(q_1|x)}{\pi_{\text{ref}}(q_1|x)} - \beta \log \frac{\pi_\theta(q_2|x)}{\pi_{\text{ref}}(q_2|x)}\Bigg)\Bigg] \\ +&\mathbb{E}_{(\tau_w,\tau_l)\sim\mathcal{D}}\Bigg[-\log\sigma\Bigg(\beta \log \frac{\pi_\theta(y_1|[x, q_1])}{\pi_{\text{ref}}(y_1|[x, q_1])} - \beta \log \frac{\pi_\theta(y_2|[x, q_2])}{\pi_{\text{ref}}(y_2|[x, q_2])}\Bigg)\Bigg], \end{aligned}$$

where $\tau_w = (x, q_1, [x, q_1], y_1)$, $\tau_l = (x, q_2, [x, q_2], y_2)$, and $(q_1, q_2, y_1, y_2) \sim \pi_{ref}$. Since λ and δ are constants, we ignore them in the derivation for simplicity. The last step in the above derivation applies Jensen's inequality to decouple the joint objective function into two terms, satisfying different preference alignments of two reward functions $r(x, q)$ and $r([x, q], y)$. Thus, by substituting the reward functions into RHS of the last derivation step, we finally get the loss function of our proposed SPO as

$$\begin{aligned} \mathcal{L}(\tau_w, \tau_l) &= \mathbb{E}_{(\tau_w,\tau_l)\sim\mathcal{D}}\left[-\log\sigma\left(r(x, q_1) - r(x, q_2)\right)\right] \\ &+ \mathbb{E}_{(\tau_w,\tau_l)\sim\mathcal{D}}\left[-\log\sigma\left(r([x, q_1], y_1) - r([x, q_2], y_2)\right)\right]. \end{aligned}$$

Table 1. We analyze the hyper-parameter settings of SFT-7B trained with SPO: 1) freeze and unfreeze the vision encoder weight during training; 2) impact of hyper-parameters including δ, and β. The first row is our final setting.

Vision Encoder	δ	β	Image		Multi-Image		Video	
			MME	LLaVA-Wilder	LLaVA^{I}	MuirBench	EgoSchema	MLVU
Freeze	1.2	0.3	1585	70.2	67.9	46.7	60.5	65.3
Freeze	1.5	0.2	1568	69.8	67.6	46.0	58.0	64.9
Freeze	1.5	0.3	1571	70.0	67.4	46.2	60.3	65.1
Unfreeze	1.2	0.2	1563	69.0	67.0	45.3	59.5	65.0

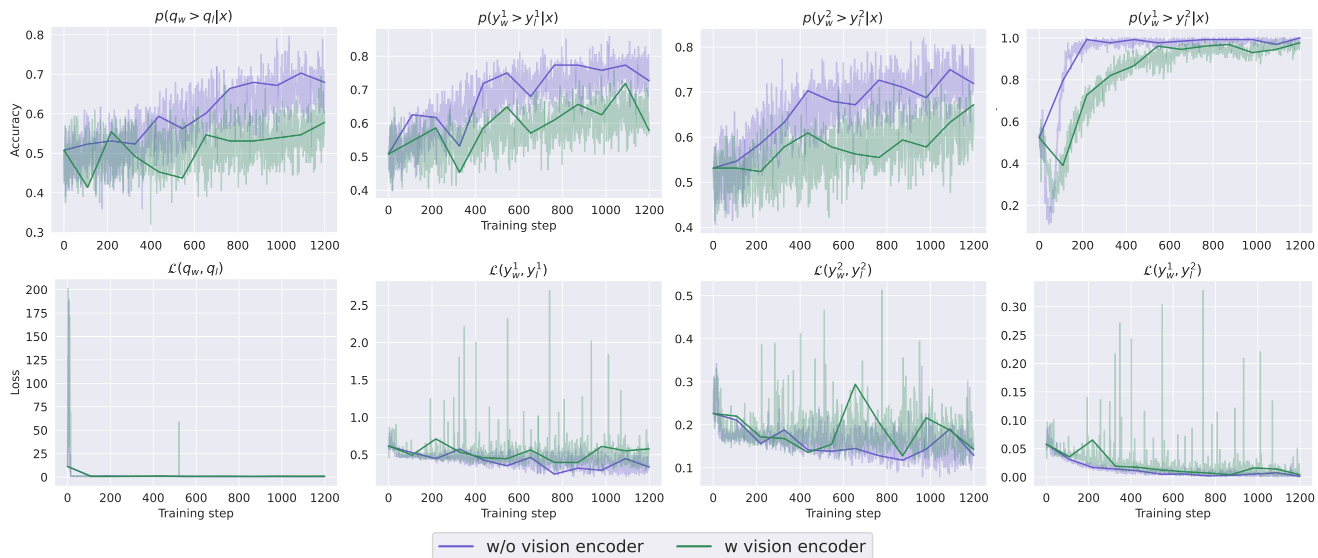


Figure 1. Preference classification accuracy and loss of SFT-7B trained with SPO during training. We compare between the models trained with and without updating vision encoder.

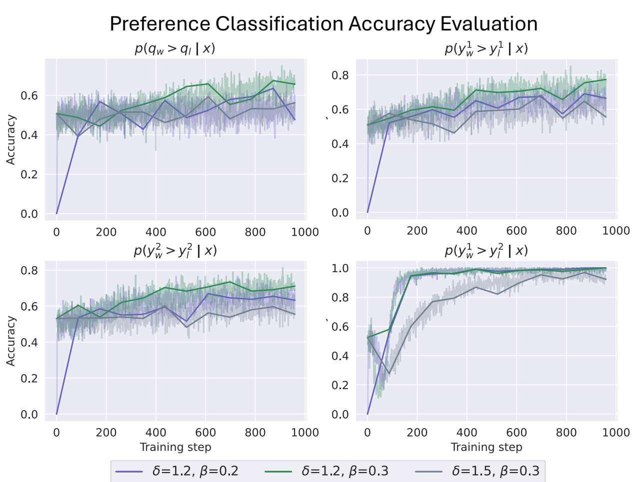


Figure 2. Preference classification accuracy of SPO trained by different settings over the course of training.

3. More results

3.1. Experimental Settings

Supervised Fine-tuning for Learning to Question. We follow SQ-LLaVA [5] by fine-tuning the our model with visual self-questioning objective function [5] on a subset of training data [3] to obtain an SFT model with questioning ability. Specifically, we modify the visual instruction data by randomly replacing the role token `[usr]` by `[vusr]` within the multi-turn conversations and maximize the log-likelihood of the `[vusr]` instructions. In this way, we can prompt the SFT model with `[vusr]` for instruction generation. For image and video pre-processing, we follow LLaVA-Onevision by leveraging the multi-scale visual embedding. During training, we optimize the trainable parameters θ using AdamW [4] with a learning rate of 2×10^{-5} and a constant scheduler for 1 epoch.

Preference Optimization. We use the SFT model from stage 1 to initialize the policy π_θ and reference model π_{ref}. Unlike the previous method [1], this work utilized the SFT model to generate 100k question-answer pairs as preference dataset. We set the sampling temperature to 1.2 for instruction and 0.75 for answer generation to ensure diversity. Then we fine-tune π_θ with a learning rate of 5×10^{-6} for 1 epoch using SPO and DPO loss on the same preference dataset. We find that tuning hyper-parameters is crucial for achieving optimal performance with all the offline preference optimization algorithms, including SPO. Gener-

Prompt GPT-4o to score each of the Instructions

```
messages = [ {"role": "User", "content": """
Text: {
[Question 1]\n … \n\n[End Question 1]\n\n
[Question 2]\n … \n\n[End Question 2]\n\n
[System] We would like to request your feedback on the quality of two questions displayed above. Given a
visual input and two corresponding questions, your task is to act as an impartial evaluator and evaluate the
questions based on the following criteria:
•  Visual Dependency: How well does the question utilize and rely on the visual input for its relevance?
•  Accuracy: How factually correct and relevant is the question in relation to the visual input?
•  Logical Flow: Does the question follow a clear and logical structure?
After evaluating, each question receives an overall score on a scale of 1 to 10. Please output a json object
containing the following information:
{
        Question 1: int //score of Question 1
        Question 2: int // score of Question 2
}}
Image:{<image>}"""}]
```

Figure 3. Prompt for GPT-4o to perform scoring on two instructions generated by our policy model.

Prompt GPT-4o to score each of the answers

```
messages = [ {"role": "User", "content": """
Text: {
[Question]\n …? \n\n
[Answer 1]\n … \n\n[End Answer 1]\n\n
[Answer 2]\n … \n\n[End Answer 2]\n\n
[System] Given a visual input, an instruction, and two corresponding answers, your task is to serve as an
impartial evaluator and score the quality of each answer. Assess the answers based on the following
criteria, aligned with human preferences:
•  Relevance to Visual Input: How accurately and effectively does the answer address the visual content?
•  Instruction Alignment: Does the answer follow the provided instruction closely and appropriately?
•  Clarity and Coherence: Is the answer easy to understand, logically structured, and free of ambiguity?
•  Human Preference: Consider how a human might prefer the answer based on helpfulness, natural tone,
   and appropriateness.
After evaluating, each question receives an overall score on a scale of 1 to 10. Please output a json object
containing the following information:
{
        Answer 1: int //score of Answer 1
        Answer 2: int // score of Answer 2
} }
Image:{<image>}"""}]
```

Figure 4. Prompt for GPT-4o to perform scoring on two answers generated by our policy model conditioned on the same visual input and instruction.

ally, we follow [6] to adaptively set $\beta = 0.5$ for pairs with larger differences (e.g., τ_1 vs τ_4), and $\beta = 0.25$ for pairs with higher similarity (τ_1 vs τ_2 and τ_3 vs τ_4).

Preference Labeling. A hybrid approach integrating LLM-as-a-judge and human annotations is a great suggestion. This work utilizes GPT-4o as a judge for preference labeling, mainly due to a trade-off between scalability and practical costs. To assess the potential bias, we evaluate the labeling score distributions of three LLMs (Fig. 5 in the manuscript) and find that GPT-4o empirically better aligns with human preference. In Fig. 3 and Fig. 4, we provide the manually designed prompt for GPT-4o for score prediction.

3.2. Model Discussion

Influence of Vision Encoder. Previous work [3] claims better performance by updating the vision encoder and LLM together during preference optimization. However, this work has a contrary observation. Fig. 1 shows that the preference classification loss becomes unstable, and the preference classification accuracy increases trivially when updating the vision encoder. The evaluation results show that the degradation of SPO w vision encoder on 4 benchmarks indicates that the visual representation may drastically be changed during preference optimization, causing over-fitting issues and damaging the model's generalizability. We find increasing the preference data scale can mitigate this issue.

Loss Term Analysis. Our final loss function (Eq.11) consists of three terms to define the desired preferences: $(q_w, y_w^1) > (q_w, y_l^1)$, $(q_w, y_w^1) > (q_l, y_l^2)$, and $(q_l, y_w^2) > (q_l, y_l^2)$. Notably, we did not include the loss term $\mathcal{L}(\tau_1, \tau_3)$: $(q_w, y_w^1) > (q_l, y_w^2)$, since $y_w^1 > y_w^2$ can be controversial as they are all preferred answers in local optima. To validate this hypothesis, Table 2 provides a comparison between 1) SPO without using $\mathcal{L}(\tau_1, \tau_3)$ and 2) with this loss term, indicating that $(q_w, y_w^1) > (q_l, y_w^2)$ may confuse the model.

Table 2. Comparison of SPO-0.5B with and without $\mathcal{L}(\tau_1, \tau_3)$

$\mathcal{L}(\tau_1, \tau_3)$	MME	LLaVA-W	MLVU	VideoMME
w/o (**ours**)	**1251**	**74.1**	**52.4**	**47.7**
with	1205	72.8	50.8	45.4

3.3. Analysis of Hyper-parameters

Hyper-parameters are crucial for achieving optimal performance with all the offline preference optimization algorithms, including SPO. Generally, we propose an adaptive β strategy to eliminate the influence of the similarity gap between the preferred and dispreferred action [6]. Specifically, the regular value of β is in range $[0.1,\ 0.5]$, and this work tests $\beta \in \{0.2, 0.3\}$ since it has overall better performance under various model scale [6]. In addition, we study the impact of δ, as it controls the importance weight of answer reward during SPO training. As shown in Eq 2.2, we simultaneously optimize preference question and answer classification. Setting $\delta > \lambda$ is straightforward since the benchmarks primarily evaluate the model's answering ability. In Table 1, we provide the evaluation results of SFT-7B trained with SPO under different hyper-parameter settings.

3.4. Qualitative Results

Fig. 5, Fig. 6, Fig. 7 and Fig. 8 show that when trained with SPO, the model's generation of a given image becomes more visually relevant than DPO. Specifically, the generation of SPO-7B has less divergent content that is nondirectly related to the image, such as "sign is often used in places like construction sites or private property" generated by DPO-7B. Additionally, when trained with DPO, the model becomes over-confident when indicating low visual dependency content (refer to Fig. 5 and Fig. 7). In contrast, SPO improves the model's visual reasoning ability and prevents over-confident prediction when explaining uncertain situations, acting similarly to the reference model, GPT-4V. The same observation can also be found in the video question answering in Fig. 9 and Fig. 10.

References

[1] Hritik Bansal, Ashima Suvarna, Gantavya Bhatt, Nanyun Peng, Kai-Wei Chang, and Aditya Grover. Comparing bad apples to good oranges: Aligning large language models via joint preference optimization. *ArXiv*, 2024. 2

[2] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017. 1

[3] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *ArXiv*, 2024. 2, 3

[4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 2

[5] Guohao Sun, Can Qin, Jiamian Wang, Zeyuan Chen, Ran Xu, and Zhiqiang Tao. Sq-llava: Self-questioning for large vision-language assistant. In *ECCV*, 2024. 2

[6] Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. β-dpo: Direct preference optimization with dynamic β. *ArXiv*, 2024. 3, 4

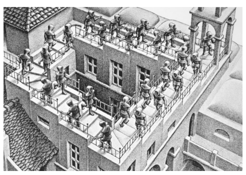
Where is the fuel cap on this car?

SPO-7B The image you've provided is of a car's fuel gauge, which is typically located on the dashboard. The fuel cap, also known as the fuel filler door, is where you would access the fuel tank to refuel your vehicle. It is usually located on the rear side of the car, either on the driver's side or the passenger's side, depending on the vehicle's design and manufacturer. To find it, look for a small door or panel with a symbol indicating where to locate the fuel cap. If you're unsure about the exact location on your specific vehicle, consult your car's owner's manual for precise instructions.

DPO-7B The fuel cap on this car is located at the left side of the dashboard, just above the center console. It's typically found near the driver's side footwell area.

GPT-4V The fuel cap on this car is located on the right side. This can be determined by the small arrow next to the fuel pump icon on the fuel gauge. The arrow points to the right, indicating the side of the vehicle where the fuel cap is located. *Reference answer*

Figure 5. Qualitative results of LLaVA-Bench-Wilder. When trained with DPO, the model becomes over-confident when indicating low visual dependency content. In contrast, SPO improves the model's visual reasoning ability and prevents over-confident prediction when explaining uncertain situations, acting similarly to the reference model, GPT-4V. We highlight the high visual relevant and low visual relevant prediction.



What is special in this image?

SPO-7B This image is a classic example of an impossible object, often referred to as an "impossible staircase" or "Penrose stairs." . It shows a structure that defies the laws of physics and perspective, creating an illusion of a three-dimensional object that cannot exist in reality. The drawing appears to depict a building with multiple levels and windows, but it's actually a trick of the eye where the lines and shading create the illusion of depth and continuity when, in fact, it's a continuous pattern that doesn't make sense in three-dimensional space.

DPO-7B The structure in the image is impossible because it violates the fundamental principles of geometry and physics. In a real-world scenario, objects cannot have multiple layers that loop back on themselves without any beginning or end. The stairs would need to be connected in a way that allows them to move seamlessly from one level to another, but this is not possible due to the constraints of three-dimensional space.

GPT-4V The image you provided is a classic example of an M.C. Escher-inspired artwork, known for its use of optical illusions and impossible constructions. This particular drawing features a series of figures continuously marching in a loop on a series of interconnected staircases on a building. The special aspect of this artwork is its depiction of a physically impossible structure, where the staircases form a never-ending loop, defying the laws of gravity and traditional architectural logic. Escher's work often plays with perspective and space to create endlessly looping paths, like the one seen here, where figures appear to ascend and descend in an eternal cycle without ever reaching a definitive end. This kind of artistic exploration challenges viewers' perceptions of reality and encourages a deeper consideration of the complexities of space and dimension. *Reference answer*

Figure 6. Qualitative results of LLaVA-Bench-Wilder.

What is funny in this image.

SPO-7B The image shows a character from the animated television series "The Simpsons." The character is standing next to a sign that reads "KEEP OUT" with an additional humorous line underneath saying "OR ENTER. I'M A SIGN, NOT A COP." This is special because it plays on the wordplay and the visual pun of the sign itself being a sign, which typically directs people not to enter, but in this case, it humorously suggests that it's not there to act as a police officer (cop) would.

DPO-7B The image features a scene from the animated television series "The Simpsons." It shows the character Lisa Simpson standing in front of a sign that reads "KEEP OUT" on top and "OR ENTER. I'M A SIGN, NOT A COP" on the bottom. The humor lies in the wordplay and the unexpected twist that the sign is not a real police officer but just a sign trying to deter people from entering an area. This type of sign is often used in places like construction sites or private property to indicate restricted access without being confused with a law enforcement officer.

GPT-4V The humor in this image comes from the contradictory messages on the sign. The sign first tells the viewer to "KEEP OUT" but then humorously contradicts itself with the lower part stating, "OR ENTER. I'M A SIGN, NOT A COP." This juxtaposition plays on the expectation that signs give authoritative, clear instructions. Instead, this sign acknowledges its own inability to enforce its commands, reducing its role to merely a passive informer, not an enforcer. This self-aware, contradictory message delivers a comical twist that highlights the limitations of signs and the absurdity of expecting them to control behavior. Such humor is typical in "The Simpsons," which often uses visual gags and wordplay to make light-hearted commentary on everyday objects and situations. *Reference answer*

Figure 7. Qualitative results of LLaVA-Bench-Wilder.

What does the smoke behind the car suggest about the driver's actions?

SPO-7B The thick, white smoke billowing from the rear of the vehicle suggests that the driver is performing a burnout. This is a driving maneuver where the driver accelerates rapidly while keeping the car stationary, causing the tires to spin and generate a lot of smoke. It's often done to warm up the tires for better traction or simply as a display of power and control. The fact that the smoke is substantial and the car is in motion indicates that this is an intentional action by the driver rather than a result of losing control.

DPO-7B The smoke behind the car suggests that the driver has recently engaged in a burnout or is in the midst of one. A burnout is a driving maneuver where the driver rapidly accelerates and then quickly releases the accelerator to spin the tires, creating a cloud of smoke as the rubber heats up and loses traction. This action is often performed for exhibition purposes, such as at car shows or enthusiast gatherings, to demonstrate the power and capabilities of the vehicle's drivetrain. The fact that there is visible smoke indicates that the tires are generating friction against the pavement, which is characteristic of a burnout.

GPT-4V The smoke behind the car in the image suggests that the driver is performing a burnout. A burnout occurs when a driver keeps a vehicle stationary and spins the wheels, causing the tires to heat up and produce smoke due to the friction with the road surface. This action is often done to warm up the tires before a race for better traction, or for show at car events to display the power and performance of the vehicle. In the context of this photo, it highlights the car's powerful engine and its capability to rapidly accelerate, indicated by the force exerted by the tires against the pavement.

Reference answer

Figure 8. Qualitative results of LLaVA-Bench-Wilder.

Explore the thematic and visual aspects of the video comprehensively.

SPO-7B The video features a football game with players wearing team uniforms. One team is in white jerseys with red accents and gold helmets, while the other team is in dark blue jerseys with white accents and dark helmets. The player in focus is wearing a white jersey with the number 65 and gold helmet. He appears to be involved in a tackle or collision with an opponent from the opposing team. The scene captures the intensity of the sport as the players engage in physical contact on the field.

DPO-7B The video contains multiple frames that capture a football game. The players are dressed in team uniforms, with one team wearing white jerseys with red and gold accents, while the other team is in dark jerseys. There are multiple players engaged in physical contact as they attempt to tackle each other. The scene shows the intensity of the sport with close proximity and physical confrontation between the players. The lighting appears to be artificial, likely from stadium lights, and the environment suggests an evening or night game.

GPT-4V The video captures a progressive series of moments on the field. Players wearing two opposing teams' uniforms are seen engaging in contact, which is typical of football gameplay. The uniforms are distinctive, with one team wearing lighter-colored jerseys with prominent numbering, while the other team's jerseys are darker. Helmets, shoulder pads, and other protective gear indicative of the sport are visible. Some frames focus on individual players, while others show a group interaction. The environment suggests this is a professional setting, likely captured from a broadcast due to the presence of a watermark indicating a television network.

Reference answer

Figure 9. Qualitative results of Video-VQA-Captioning.

Elaborate on the visual and narrative elements of the video in detail.

SPO-7B The video consists of two clips. In the first clip, an individual is seen celebrating in a sports arena with bright lights and a crowd in the background. The person is wearing a white jersey and appears to be making a victory gesture. The scoreboard visible in the background shows a basketball game with scores for the teams "LAC" and ”MIN”. The second clip features a similar scene with a different individual in a celebratory pose, also with a basketball game scoreboard in the background displaying the same teams' scores.

DPO-7B The video contains a series of images taken from a live sports event, showing an individual in a white jersey with the number 17 on it. The setting appears to be an indoor stadium filled with spectators. In one image, the individual is seen celebrating with fans around him. The scoreboard visible in the background indicates that the game is a basketball match between two teams, with scores and time remaining in the fourth quarter. The text overlay provides context about the event, mentioning the Super Bowl and an incident involving money owed to a Philadelphia mob.

GPT-4V The video consists of multiple identical frames showing what appears to be a basketball game celebration. Each frame captures the audience in the background, cheering and standing up, while the foreground focuses on a professional basketball player who is celebrating, likely after a successful play or game victory. The player is wearing a white headband, a white tank top with blue and black trim, and shorts. It's difficult to discern the exact details of the venue, but it resembles a professional basketball arena. There's also an overlay of a basketball game's scoreboard at the bottom of the frames, indicating teams and their respective scores. However, the focus is on the player's reaction and the crowd's excitement. *Reference answer*

Figure 10. Qualitative results of Video-VQA-Captioning.