

## Prompt Exploration

### MODÈLES

Les modèles testés sont les suivants :

- llava:7b (7b)
- llava:13b (13b)
- llava:34b (34b)
- llava-llama3 (8b)
- llama3.2-vision (11b)
- gemma3:4b (4b)
- gemma3:12b (12b)
- gemma3:27b (27b)

Les résultats observés sur 13 prompts sont les suivants :

- llama3.2-vision a obtenu les meilleurs scores pour chaque prompt testé
- les modèles llava ont tendances à obtenir les pires résultats en moyenne

gemma3:12b	gemma3:27b	gemma3:4b	llama3.2-vision	llava	llava-llama3	llava:13b	llava:34b	
31	30	29	40	25	22	28	20	
26	27	18	33	19	17	21	14	
28	30	24	35	25	21	25	25	
26	33	21	30	25	24	25	17	
29	33	23	39	25	26	26	22	
18	11	13	23	12	11	11	11	
22	22	23	32	17	22	20	23	
27	32	24	37	24	13	25	22	
29	31	27	36	27	22	25	26	
34	35	28	38	17	26	23	21	
36	35	27	36	24	24	26	25	
30	29	25	37	26	23	26	24	
31	33	24	33	22	27	25	24	
367	381	306	449	288	278	306	274	TOTAL
3	2	4	1	6	7	4	8	
Max	Max-1	Max-2	Max-3-5		Min+3-5	Min+2	Min+1	Min

- la taille des modèles semble avoir un impact assez imprévisible, par exemple pour llava on a eu les résultats suivants llava:34b < llava:7b < llava:13b, donc le plus grand modèle a eu les pires résultats et le modèle de taille moyenne a eu les meilleurs résultats.

Pour gemma3 les scores ont suivi la taille du modèle.

Mais les tailles peuvent réussir/rater sur différentes images, par exemple :

[illegible]

2.png

[illegible]

7.png

- La plupart du temps gemma3:12b réussit à identifier correctement 2.png mais pas 7.png alors que gemma3:27b réussit à identifier correctement 7.png mais pas 2.png.

llava et gemma réussissent/ratent sur différentes images, par exemple :

[illegible]

04.png

[illegible]

39.png

- La plupart du temps les modèles llava réussissent à identifier 04.png mais pas 39.png alors que gemma3 réussissent à identifier 39.png mais pas 04.png.

Les réponses longues ont tendance à obtenir des résultats un petit peu mieux en moyenne d'après notre méthode d'évaluation (test si un des termes correct est présent dans la réponse), mais cela est au prix des conséquences suivantes :

- À cause de ces raisons nous ne comptons pas tester plus de prompts avec des réponses longues.

Si un bras robot tient l'objet à identifier et prend une partie significative de l'image, il vaut mieux indiquer dans le prompt qu'il faut ignorer le bras robot ou donner un scénario dans lequel il est normal que le bras soit présent afin d'éviter que le modèle essaye d'identifier le bras au lieu de l'objet.

Il est également possible de donner un contexte dans lequel il est normal d’avoir un bras robot tenant un objet, par exemple le prompt “Act as a robotic assistant that identifies tools. What is in your grip?”

prompt	gemma3.12b	gemma3.27b	gemma3.4b	llama3.2-vision	llava	llava-llama3	llava-13b	llava-34b	TOTAL
Based on what you see, you must identify the object. Be concise, identify the object.	26	27	18	33	19	17	21	14	175
Based on what you see, you must identify the object. You will see a robotic arm in the image, but IGNORE IT, focus on the object. Be concise, identify the object.	28	30	24	35	25	21	25	25	213
You are a robot. Based on what you see, you must identify the object. Be concise, identify the object.	27	32	24	37	24	13	25	22	204
You are a robot. Based on what you see, you must identify the object. You will see a robotic arm in the image, but IGNORE IT, focus on the object. Be concise, identify the object.	29	31	27	36	27	22	25	26	223

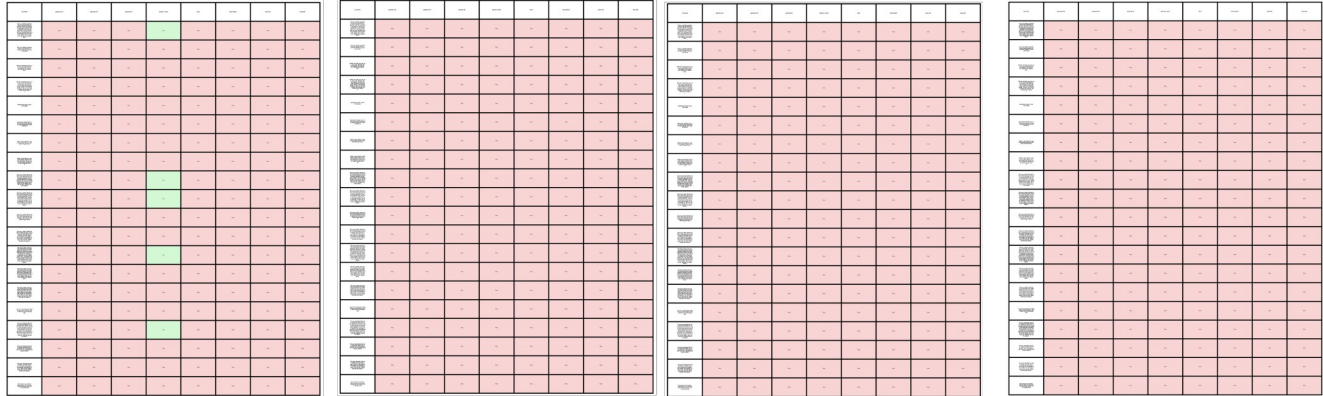
Pour les images sans bras robotique (images 48 à 60) cette mention erronée ne semble pas avoir d'impact.

Les modèles ont souvent tendance à détecter correctement/incorrectement toutes images d'un même objet, par exemple

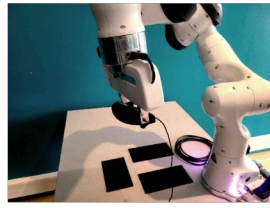
```
llama3.2-vision | 111111111111000000001111001011110000000000001001111111111111110 total : 33/27/60
                  mug spo. glo. pla. fan soa. con. clo. mo. scr. pcb cle. ....
```

ici llama3.2-vision détecte correctement toutes les images de mug et aucune des images de l'assiette.

certaines images ne sont jamais identifiées correctement, par exemple presque toutes les images de  
 de bras tenues par le bras robot ne sont jamais reconnues



33.png



34.png



35.png



36.png

Parfois ils détectent que c'est un objet tenu par le bras mais n'arrivent pas à l'identifier, parfois ils pensent que c'est une partie du robot (probablement à cause de la couleur)

prompt : "Ignore the robotic arm. Focus only on the object being held. Identify it."

```
model : llama3.2-vision
```

image : 34.jpg

réponse : "The **object being held by the robotic arm appears to be a small, rectangular piece of black material**. It is likely a piece of plastic or metal, possibly used for testing or demonstration purposes. The object's shape and size suggest that it may be a component of a larger device or system, but without more information, it is difficult to determine its exact purpose or function."

prompt : “You are a robot. You are analyzing this image to detect the item held. You will see a robotic arm in the image, but IGNORE IT, focus on the object. Be concise, identify the object.”

```
model : llava:7b
```

image : 37.png

réponse : "The object in the image is a robotic arm **with a camera mounted on it.**"

Les 3 rectangles visibles sur la table posent parfois problèmes car le modèle commence à les décrire, par exemple :

prompt : “You are a robot. Based on what you see, you must identify the object. Be concise, identify the object.”

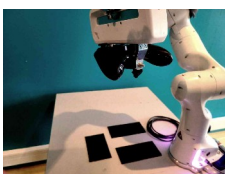


image : 28.jpg

modèle : llama3.2-vision

réponse : “The object in the image appears to be a collection of black cards or rectangles, possibly used for educational or artistic purposes. The cards are arranged on a flat surface and have a uniform size and shape. They may be made of cardboard, paper, or another material. Without more context, it is difficult to determine their exact purpose or meaning.”

## MEILLEURS PROMPTS TROUVES

En se basant sur les meilleurs prompts trouvés plus haut, on a testé quelques combinaisons des parties qui avaient l’air de fonctionner.

Les 7 prompts testés ont les meilleurs scores trouvés jusqu’à présent, ils ont tous des scores assez proches.

	gemma3:12b	gemma3:27b	gemma3:4b	llama3.2-vision	llava	llava-llama3	llava:13b	llava:34b		
You are a robot. You are analyzing this image to detect the item held. You will see a robotic arm in the image, but IGNORE IT, focus on the object. Be concise, identify the object.	33	34	28	40	28	25	27	30	245	3
Act as a robotic assistant that identifies tools. Based on what you see, you must identify the object gripped in your robotic arm. Disregard the arm and background clutter, focus on the object. Be concise, identify the object.	32	34	25	40	29	27	26	28	241	6
You are analyzing this image to detect the item held. Based on what you see, you must identify the object gripped in your robotic arm. Disregard the arm and background clutter, focus on the object. Be concise, identify the object.	35	33	29	41	27	24	26	28	243	5
You are a robot. Based on what you see, you must identify the object gripped in your robotic arm. Disregard the arm and background clutter, focus on the object. Be concise, identify the object.	36	36	28	40	27	27	26	28	248	1
You are a robot. You are analyzing this image to detect the item held. Based on what you see, you must identify the object gripped in your robotic arm. Disregard the arm and background clutter, focus on the object. Be concise, identify the object.	34	33	27	41	27	24	26	26	238	7
You are a robot. You are analyzing this image to detect the item held. Disregard the arm and background clutter, focus on the object. Be concise, identify the object.	34	34	31	37	28	26	28	30	246	1
You are a robot. Based on what you see, you must identify the object gripped in your robotic arm. You will see a robotic arm in the image, but IGNORE IT, focus on the object. Be concise, identify the object.	35	34	26	42	28	25	27	28	245	3
	239	238	194	281	194	178	186	198	TOTAL	
	2	3	5	1	5	8	7	4		RANK
	Max	Max-1	Max-2	Max-3	Max-4	Max-5	Max-6	Max-7		

Le prompt suivant est le meilleur retenu pour la suite des tests :

“You are a robot. Based on what you see, you must identify the object gripped in your robotic arm. Disregard the arm and background clutter, focus on the object. Be concise, identify the object.”

## RÉSULTATS SUR LES 20 PROMPTS

	gemma3:12b	gemma3:27b	gemma3:4b	llama3.2-vision	llava	llava-llama3	llava:13b	llava:34b		
Act as a robotic assistant that identifies tools. Based on what you see, you must identify the object gripped in your robotic arm. Disregard the arm and background clutter, focus on the object. Be concise, identify the object.	32	34	25	40	29	27	25	28	241	6
Act as a robotic assistant that identifies tools. What is in your grip?	31	30	29	40	25	22	28	20	225	9
Based on what you see, you must identify the object. Be concise, identify the object.	26	27	18	33	19	17	21	14	175	19
Based on what you see, you must identify the object. You will see a robotic arm in the image, but IGNORE IT, focus on the object. Be concise, identify the object.	28	30	24	35	25	21	25	25	213	15
Identify the item in the image	26	33	21	30	25	24	25	17	201	17
Ignore the robotic arm. Focus only on the object being held. Identify it.	29	33	23	39	25	26	26	22	223	11
What is the object in the image? Answer with a single word.	18	11	13	23	12	11	11	11	110	20
What is the object in the image? Ignore the robot arm, focus on the object, Answer with a single word.	22	22	23	32	17	22	20	23	181	18
You are a robot. Based on what you see, you must identify the object gripped in your robotic arm. Disregard the arm and background clutter, focus on the object. Be concise, identify the object.	36	36	28	40	27	27	26	28	248	1
You are a robot. Based on what you see, you must identify the object gripped in your robotic arm. You will see a robotic arm in the image, but IGNORE IT, focus on the object. Be concise, identify the object.	35	34	26	42	28	25	27	28	245	3
You are a robot. Based on what you see, you must identify the object. Be concise, identify the object.	27	32	24	37	24	13	25	22	204	16
You are a robot. Based on what you see, you must identify the object. You will see a robotic arm in the image, but IGNORE IT, focus on the object. Be concise, identify the object.	29	31	27	36	27	22	25	26	223	11
You are a robot. You are analyzing this image to detect the item held. Based on what you see, you must identify the object gripped in your robotic arm. Disregard the arm and background clutter, focus on the object. Be concise, identify the object.	34	33	27	41	27	24	26	26	238	7
You are a robot. You are analyzing this image to detect the item held. Disregard the arm and background clutter, focus on the object. Be concise, identify the object.	34	34	31	37	28	26	28	30	248	1
You are a robot. You are analyzing this image to detect the item held. You will see a robotic arm in the image, but IGNORE IT, focus on the object. Be concise, identify the object.	33	34	28	40	28	25	27	30	245	3
You are a warehouse robot sorting objects. What are you holding?	34	35	28	38	17	26	23	21	222	12
You are analyzing this image to detect the item held. Based on what you see, you must identify the object gripped in your robotic arm. Disregard the arm and background clutter, focus on the object. Be concise, identify the object.	35	33	29	41	27	24	26	28	243	5
You are analyzing this image to detect the item held. Disregard the arm or background clutter.	36	35	27	36	24	24	26	25	233	8
You must identify the object. You will see a robotic arm in the image, but IGNORE IT, focus on the object. Be concise, identify the object.	30	29	25	37	26	23	26	24	220	13
Your task is to classify the item. Do not describe the robot or background.	31	33	24	33	22	27	25	24	219	14
	606	619	500	730	482	456	492	472	TOTAL	
	3	2	4	1	6	8	5	7		RANK
	Max	Max-1	Max-2	Max-3-5		Min+3-5	Min+2	Min+1	Min	

# IMPACT DES TRANSFORMATIONS SUR LES IMAGES

## TRANSFORMATIONS

Nous avons également effectué une série de transformations sur toutes les images pour voir leurs impact sur les résultats.

Liste des transformations :



- normal (ressources/images\_base) : images de base testées jusqu'à maintenant



- haute qualité (ressources/high\_quality) : images avec la meilleur résolution possible (images capturées par le téléphone avant de faire les réductions initiales)



- bon rognage (ressources/good\_crop) : images rognées manuellement pour contenir le moins possible d'éléments autres que l'objet



- rognage coupant (cut\_object\_crop) : images rognées manuellement pour l'image ne contienne que une partie de l'objet, l'objet n'est également pas centré dans l'image



- flou (ressources/blurred) : images floutées



- compression (ressources/jpeg\_artifacts) : images enregistrées avec un fort taux de compression



- bruit impulsionnel (ressources/salt\_pepper\_noise) : images avec du bruit impulsionnel



- retournement horizontal (ressources/flipped\_horizontal) : images retournées horizontalement (effet miroir gauche-droite)



- retournement vertical (ressources/flipped\_vertical) : images retournées verticalement (effet miroir haut-bas)



- rotation (ressources/rotated) : images pivotées d'un certain angle (45 degrés).





- niveaux de gris (ressources/grayscale) : images converties en noir et blanc



- teinte bleue (ressources/tint\_blue) : images avec une dominante bleue ajoutée artificiellement



- teinte verte (ressources/tint\_green) : images avec une dominante verte ajoutée artificiellement



- teinte rouge (ressources/tint\_red) : images avec une dominante rouge ajoutée artificiellement



- surexposition (ressources/overexposed) : images artificiellement éclaircies pour simuler une exposition trop forte



- sous-exposition (ressources/underexposed) : images artificiellement assombries pour simuler un manque de lumière



- occlusion partielle (ressources/occluded) : images où des parties de l'objet sont recouvertes par des rectangles noirs simulant des obstructions (3 rectangles aléatoires).  
Notes : il aurait été mieux de créer les rectangles manuellement pour chaque images car les positions générées aléatoirement sont parfois très petites et dans des zones peu gênantes.

D'autant plus que nos images ont déjà des rectangles noirs sur la table par défaut que on ne pouvait pas enlever.



## RÉSULTATS

Table du nombre d'images correctement identifiées par couple modèle/transformation sur 60.

	gemma3:12b	gemma3:27b	gemma3:4b	llama3.2-vision	llava	llava-llama3	llava:13b	llava:34b	TOTAL
blurred	22	17	15	30	15	18	22	24	163
cut_object_crop	26	29	13	28	34	20	33	35	218
flipped_horizontal	34	32	25	38	24	22	28	27	230
flipped_vertical	27	29	24	41	21	16	22	25	205
good_crop	37	33	33	38	36	31	38	37	283
grayscale	34	34	24	39	19	22	22	20	214
high_quality	40	40	30	49	26	28	35	31	279
images_base	36	36	28	40	27	27	26	27	247
jpeg_artifacts	29	27	21	38	23	25	28	29	220
occluded	30	33	24	37	20	23	25	25	217
overexposed	29	24	27	39	24	23	26	28	220
rotated	28	28	18	39	14	16	20	24	187
salt_pepper_noise	24	25	19	33	14	24	25	26	190
tint_blue	35	30	24	41	27	24	24	27	232
tint_green	32	33	24	41	25	27	24	29	235
tint_red	32	31	23	39	28	24	28	26	231
underexposed	31	33	24	39	25	27	27	26	232
TOTAL	526	514	396	649	402	397	453	466	

Comparaisons par rapport aux résultats avec les images par défaut.

	gemma3:12b	gemma3:27b	gemma3:4b	llama3.2-vision	llava	llava-llama3	llava:13b	llava:34b	TOTAL
blurred	-14	-19	-13	-10	-12	-9	-4	-3	-84
cut_object_crop	-10	-7	-15	-12	7	-7	7	8	-29
flipped_horizontal	-2	-4	-3	-2	-3	-5	2	0	-17
flipped_vertical	-9	-7	-4	1	-6	-11	-4	-2	-42
good_crop	1	-3	5	-2	9	4	12	10	36
grayscale	-2	-2	-4	-1	-8	-5	-4	-7	-33
high_quality	4	4	2	9	-1	1	9	4	32
images_base	0	0	0	0	0	0	0	0	0
jpeg_artifacts	-7	-9	-7	-2	-4	-2	2	2	-27
occluded	-6	-3	-4	-3	-7	-4	-1	-2	-30
overexposed	-7	-12	-1	-1	-3	-4	0	1	-27
rotated	-8	-8	-10	-1	-13	-11	-6	-3	-60
salt_pepper_noise	-12	-11	-9	-7	-13	-3	-1	-1	-57
tint_blue	-1	-6	-4	1	0	-3	-2	0	-15
tint_green	-4	-3	-4	1	-2	0	-2	2	-12
tint_red	-4	-5	-5	-1	1	-3	2	-1	-16
underexposed	-5	-3	-4	-1	-2	0	1	-1	-15
TOTAL	-86	-98	-80	-31	-57	-62	11	7	
	10 or more	7 to 9	4 to 6	1 to 3	0 -1 to -3	-4 to -6	-7 to -9	-10 or less	

## OBSERVATIONS

- On observe que les transformations de couleur (niveaux de gris, teinte bleue, teinte verte, teinte rouge, surexposition et sous-exposition) ont eu des impacts négatif généralement plutôt faible sur les résultats.
- La qualité de l'image à un impact prévisible sur l'image, les images de haute qualité ont eu des meilleurs résultats que les images de bases (qui ont eu une étape initiale pour diminuer leur qualité pour simuler une caméra d'un robot plutôt que l'appareil photo du téléphone).
- De manière prévisible, les images bruitées (flou, bruit impulsionnel et compression) impactent négativement les résultats, cependant les résultats sont très hétérogènes, les grands modèles llava semblent être peu impactés, alors que les modèles gemma3 semblent très fortement impactés.

- Pour les transformations géométriques : Retournement horizontal ne semble pas avoir beaucoup d'effet. Retournement vertical réduit les scores bien plus, probablement car les images à l'envers sont bien plus rares. Rotation a eu un très grand impact, mais c'est peut être dû à la manière dont les images ont été créées, les contours noirs hors de l'image ainsi que les opérations pixel à pixels ont peut être impacté la qualité de l'image plus que la rotation elle même.
- Pour les rognages : de manière prévisible un bon rognage autour de l'image a tendance à augmenter les scores, en particulier pour les modèles llava. Tandis que pour un rognage mal centré (une partie de l'objet n'est pas visible, et il n'est pas centré sur l'image) les résultats sont intéressants : pour les modèles gemma3 et llama les scores ont beaucoup diminué, par contre pour les modèles llava ont eu des meilleurs résultats, probablement car l'image est un peu mieux zoomée sur l'objet malgré les autres défauts.
- Donc généralement on observe que les modèles llava semblent bien moins affectés par les transformations négatives, du moins pour les grands modèles (13b et 34b) que les modèles gemma3. Pour les transformations positives, llava est également bien meilleur pour les exploiter.
- Note importante : il faut se rappeler que généralement les modèles gemma3 ont eu des meilleurs résultats que les modèles llava pour les images de bases, et c'est toujours le cas pour la plupart des transformations ici aussi, mais les différences entre les deux sont plus petites car llava semble mieux gérer les transformations sur les images que gemma3.