UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the bachelor thesis:

Izquierdo Pérez, Marcel. *Are AI models fooled by 3D visual illusions?*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Enginyeria de Dades)

This version is available at https://ddd.uab.cat/record/337333

under the terms of the  license.

Anàlisi de la percepció d'il·lusions òptiques per models multimodals (VLMs): exactitud, estabilitat i tipologia d'errors

Marcel Izquierdo Perez
1676290@uab.cat
Curs 2025/26 · Grau en Enginyeria de Dades · Escola d'Enginyeria (UAB)

Resum*—Les il·lusions òptiques són estímuls que generen una percepció sistemàticament diferent de la realitat física i, per això, ofereixen una prova exigent per distingir si un sistema de visió interpreta una escena o si només hi reconeix patrons. Aquest treball investiga com tres models de visió-llenguatge (VLMs) de codi obert —Qwen2.5-VL-7B, InternVL2.5-8B i Phi-3.5-vision— perceben sis categories d'il·lusions òptiques. Es dissenya un banc de 150 preguntes d'opció múltiple (30 imatges × 5 variants de prompt) amb un protocol blind que barreja la posició de les opcions entre variants. L'avaluació es fa amb tres famílies de mètriques: exactitud (global i per categoria), estabilitat davant la reformulació (consistència, volatilitat i entropia) i tipologia d'errors (il·lusoris vs. distractors). Els resultats mostren que el millor model (Qwen2.5-VL, 70,7%) queda molt per sota del rendiment humà, que la volatilitat és del 100% en tots tres models —cap manté la mateixa resposta a les cinc variants de cap imatge— i que, quan els models s'equivoquen, ho fan majoritàriament triant l'opció il·lusòria (fins al 97,7% dels errors). Aquestes troballes indiquen que els VLMs actuals no posseeixen una comprensió estable de les il·lusions òptiques i que el seu encert aparent pot ser un artefacte del fraseig i de la posició de les opcions.*

Paraules clau*—Il·lusions òptiques, models de visió-llenguatge (VLM), percepció visual, estabilitat de respostes, volatilitat, errors il·lusoris, avaluació multimodal.*

Abstract*—Optical illusions are stimuli that elicit a perception systematically different from physical reality, which makes them a demanding test for distinguishing whether a vision system interprets a scene or merely recognizes patterns. This work investigates how three open-source vision-language models (VLMs) —Qwen2.5-VL-7B, InternVL2.5-8B and Phi-3.5-vision— perceive six categories of optical illusions. We design a benchmark of 150 multiple-choice questions (30 images × 5 prompt variants) using a blind protocol that shuffles option positions across variants. Evaluation relies on three metric families: accuracy (global and per-category), stability under rephrasing (consistency, volatility and entropy), and error typology (illusory vs. distractor errors). Results show that the best model (Qwen2.5-VL, 70.7%) falls well below human performance, that volatility reaches 100% for all three models —none maintains the same answer across all five variants of any image— and that, when models err, they overwhelmingly choose the illusory option (up to 97.7% of errors). These findings indicate that current VLMs lack a stable understanding of optical illusions and that their apparent accuracy may be an artifact of phrasing and option positioning.*

Index Terms*—Optical illusions, vision-language models (VLM), visual perception, response stability, volatility, illusory errors, multimodal evaluation.*

I. INTRODUCCIÓ

1.1 Context i motivació

Hi ha imatges que a primera vista semblen inofensives: una escala, una torre, una superfície amb ombres o un simple patró geomètric. Però quan l'observador s'hi atura uns segons, la imatge deixa de ser estable. Allò que semblava sortir cap enfora s'enfonsa, allò que semblava recte es torça, i allò que semblava possible es converteix en un objecte que no pot existir. Aquest petit desajust entre el que hi ha i el que sembla és, precisament, el punt de partida d'aquest treball.

En els darrers anys, els models de visió-llenguatge (VLMs) han experimentat un creixement espectacular. Arquitectures com GPT-4V [11], Gemini [12] o LLaVA [13] integren percepció visual i generació de text en un sol sistema, i responen preguntes sobre imatges, descriuen escenes i raonen sobre contingut visual amb una solvència sorprenent. Aquest èxit, però, fa més urgent una pregunta de fons: aquests sistemes realment veuen, o només reconeixen patrons estadístics? El rendiment dels VLMs davant d'estímuls dissenyats per enganyar la percepció —les il·lusions òptiques— resta encara poc explorat, i és justament aquí on la distinció entre veure i reconèixer es fa visible.

Les il·lusions òptiques són imatges que provoquen una percepció diferent de la realitat física de l'estímul [14]. Constitueixen un banc de proves privilegiat per avaluar la percepció visual perquè la resposta «correcta» (la realitat física) i la resposta «humana» (la percepció il·lusòria) estan ben definides i són sistemàticament diferents. Obliguen el model a anar més enllà del reconeixement superficial de patrons i a interpretar estructures que el sistema perceptiu humà processa de manera involuntària i sovint errònia. Per això permeten una lectura especialment rica del comportament del model: si encerta, demostra que pot superar la percepció superficial; si falla cap a l'opció il·lusòria, indica que «veu» com un humà; i si falla cap a una opció distractora, suggereix que no ha entès l'escena.

La pregunta que guia aquest treball és, doncs, senzilla d'enunciar però difícil de respondre: què veu realment un VLM quan la imatge ha estat construïda per enganyar la percepció?

1.2 Objectius

Objectiu principal: avaluar en quina mesura tres VLMs de codi obert perceben les il·lusions òptiques com un humà, com una màquina «literal» o de manera inconsistent. D'aquest objectiu general se'n deriven quatre objectius específics:

1) Dissenyar un banc d'avaluació de 30 imatges × 5 variants de prompt amb protocol blind (barreja de posicions de les opcions).

2) Mesurar l'exactitud global i per categoria de cada model.

3) Quantificar l'estabilitat de les respostes davant reformulacions equivalents del prompt.

4) Classificar els errors en il·lusoris, distractors i altres per entendre la naturalesa dels errors de cada model.

1.3 Organització del document

La secció II revisa l'estat de l'art. La secció III descriu la metodologia: il·lusions, models, prompts, protocol i mètriques. La

secció IV presenta els resultats; la secció V, la discussió; la secció VI, les conclusions i línies futures; i la secció VII, les limitacions.

II. ESTAT DE L'ART

L'estudi de les il·lusions òptiques dins la intel·ligència artificial ha passat, en pocs anys, d'anàlisis aïllades sobre casos concrets a propostes més sistemàtiques que intenten mesurar fins on arriba la visió dels models. Aquest interès creixent no és casual: a mesura que els VLMs s'han consolidat com a eines capaces de descriure escenes complexes, la pregunta sobre si realment veuen o simplement reconeixen patrons s'ha tornat més pertinent. Les il·lusions, construïdes precisament per subvertir la percepció, ofereixen un terreny on aquesta distinció es fa observable.

2.1 IllusionVQA

IllusionVQA [1] és el treball de referència més directe. Presenta dos conjunts de dades —IllusionVQA-Comprehension (435 instàncies, 12 categories) i IllusionVQA-Soft-Localization (1.000 instàncies)— per avaluar la comprensió d'il·lusions per VLMs en format d'opció múltiple. GPT-4V hi assoleix un 62,99% (4-shot) davant el 91,03% humà en comprensió, i un 49,7% davant el 100% humà en localització. Un resultat clau és que les tècniques de prompting avançat (ICL, CoT) degraden el rendiment de Gemini-Pro, cosa que evidencia una fragilitat sistèmica davant d'estímuls il·lusoris.

2.2 Is CLIP Fooled by Optical Illusions?

Ngo et al. [2] avaluen si CLIP percep onze il·lusions òptiques com els humans, usant parells de prompts (il·lusori vs. literal) i mesurant les puntuacions CLIP. Troben que CLIP és enganyat de manera fiable per 3 de les 11 il·lusions (Vasarely, contrast simultani i triangle de Kanizsa). Que CLIP només s'enganyi amb algunes de les il·lusions, i no amb totes, indica que la seva sensibilitat perceptiva varia segons l'estímul i no constitueix una propietat uniforme del model. El seu mètode multimodal (imatge + text) per sondejar la percepció és una inspiració directa per al nostre disseny.

2.3 Visual Anagrams i Illusion3D

Geng et al. (2024) [3] sintetitzen imatges que canvien d'aparença sota transformacions geomètriques, usant DeepFloyd IF amb dessorollat paral·lel. Feng et al. (2024) [4] estenen la idea a 3D amb Variational Score Distillation. Tots dos treballs se centren en la generació d'il·lusions, no en la percepció, però demostren que les representacions internes dels models de difusió capturen propietats perceptives no trivials, i introdueixen conceptes útils com l'estabilitat entre vistes i la consistència entre interpretacions.

2.4 Posicionament del nostre treball

A diferència d'IllusionVQA, que avalua models propietaris (GPT-4V, Gemini-Pro) amb una sola formulació per pregunta, aquest treball: (1) usa models de codi obert executats en local; (2) introdueix cinc variants de prompt per imatge per mesurar l'estabilitat; (3) aplica un protocol blind amb barreja de posicions; i (4) classifica els errors en il·lusoris i distractors. L'objectiu no és construir un nou benchmark exhaustiu, sinó observar amb prou detall on i per què els models encerten o fallen, de manera que es pugui respondre no només «quant encerta el model?» sinó «encerta de manera estable i per les raons correctes?».

III. METODOLOGIA

3.1 Il·lusions escollides

S'han seleccionat sis categories amb cinc imatges cadascuna (30 imatges en total), procedents principalment de la col·lecció d'Akiyoshi Kitaoka [5]. Cada categoria aïlla un mecanisme perceptiu diferent (Taula 1), de manera que el conjunt manté varietat perceptiva sense tornar-se heterogeni.

Taula 1—*Categories d'il·lusions seleccionades.*

Codi	Categoria	Efecte perceptiu
A	Ambiguous depth	Bistabilitat / reversió de profunditat
B	Bulging	Superfície convexa/còncava aparent
C	Crater illusion	Canvi convex-còncau per ombrejat
I	Impossible figures	Geometria 3D impossible
R	Reverspective	Profunditat invertida (còncau → convex)
L	Leaning Tower	Divergència aparent entre imatges idèntiques

3.2 Models

La selecció final de models va quedar condicionada per l'entorn d'execució (una màquina virtual o virtual machine, VM, amb driver CUDA 12.4 i vLLM 0.8.0). Els models inicialment previstos (Gemma 3n, Qwen3.5, GLM-4.6V) es van descartar per incompatibilitats d'arquitectura o per manca de suport de visió a vLLM. El criteri de tria és també metodològic: és preferible un experiment completament reproduïble amb models realment executables que no pas una selecció més ambiciosa sobre el paper però difícil de replicar. Els tres models finals seleccionats són Qwen2.5-VL-7B-Instruct [6], InternVL2.5-8B [7] i Phi-3.5-vision-instruct [8] (Taula 2).

Taula 2—*Models seleccionats.*

Àlies	Repo Hugging Face	Mida	Família
qwen25vl [6]	Qwen/Qwen2.5-VL-7B-Instruct	7B	Alibaba
internvl [7]	OpenGVLab/InternVL2.5-8B	8B	OpenGVLab
phi35v [8]	microsoft/Phi-3.5-vision-instruct	4,2B	Microsoft

Entorn: VM VM_IA_03 amb GPU NVIDIA, CUDA 12.4, vLLM 0.8.0, PyTorch 2.6.0+cu124, transformers 4.49.0, Python 3.12, entorn conda vllm124.

3.3 Disseny de prompts i protocol blind

El disseny de les preguntes s'inspira en el protocol de Ngo et al. [2], que sonda la percepció dels models contraposant, per a cada estímul, una opció literal (la realitat física) i una opció il·lusòria (la percepció humana). Per a cada imatge s'han creat cinc variants (V0–V4) del prompt (Taula 3). Les opcions de resposta es mantenen semànticament fixes i només es modifica l'enunciat, de manera que, si la resposta del model canvia, el canvi es pugui atribuir a la formulació i no a una alteració de les opcions. Cada pregunta té quatre opcions (A/B/C/D), cadascuna etiquetada internament com a literal (correcta), il·lusòria (percepció humana) o distractora. Crucialment, entre variants d'una mateixa imatge les opcions canvien de posició (protocol blind), de manera que cap model pugui encertar per biaix posicional. Tots els prompts són en català.

Taula 3—*Variants de prompt.*

Variant	Transformació
V0	Base
V1	Sinònims perceptius
V2	Reformulació indirecta
V3	Canvi de registre (col·loquial)
V4	Marc instruccional («Observa bé…»)

3.4 Protocol d'execució

Taula 4—*Paràmetres d'execució.*

Paràmetre	Valor	Justificació
Temperatura	0.0 (greedy)	Determinisme; la variació ve del prompt
max_tokens	10	Només cal una lletra (A/B/C/D)
System prompt	«Answer ONLY with the letter…»	Resposta atòmica
Crides/model	150 (30 img × 5 var)	Banc complet
Total crides	450 (3 × 150)	—

Tots els models reben exactament les mateixes 150 preguntes, amb el mateix text de sistema i els mateixos paràmetres de generació, de manera que qualsevol diferència observada sigui atribuïble al model i no a la manera de preguntar.

3.5 Mètriques d'avaluació

Més enllà de comptar quantes respostes són correctes, el treball vol entendre com i per què els models encerten o fallen. Per això les mètriques s'organitzen en tres famílies que responen a tres preguntes diferents: si el model encerta, si manté la resposta quan la pregunta es reformula i, quan s'equivoca, quin tipus d'error comet.

Per a cada imatge es defineix una resposta literal —la que descriu la realitat física de l'estímul (per exemple, que dues línies fan la mateixa llargada)— i una resposta il·lusòria —la que descriu la percepció humana provocada per la il·lusió (per exemple, que una línia sembla més llarga que l'altra).

Família 1 — Exactitud: exactitud global i per categoria (proporció de respostes correctes) i taxa de lectura il·lusòria (proporció de respostes que cauen a l'opció il·lusòria).

Família 2 — Estabilitat: fracció majoritària (proporció de variants que coincideixen amb la resposta més freqüent per imatge), entropia de Shannon de la distribució de respostes i taxa de volatilitat (% d'imatges on la resposta canvia entre variants).

L'entropia de Shannon [15] de la distribució de respostes d'una imatge es calcula com $H = -\Sigma_i\ p_i \cdot \log_2(p_i)$, on p_i és la proporció de variants que han triat l'opció i ($i \in \{A, B, C, D\}$). Una entropia de 0 bits indica una resposta perfectament estable; el màxim de 2 bits (amb quatre opcions) indica una distribució uniforme, és a dir, màxima inestabilitat.

Família 3 — Tipologia d'errors: per a cada error, es classifica si la resposta cau a l'opció il·lusòria o a una distractora, distingint així els «errors humans» (il·lusoris) dels «errors per confusió» (distractors).

IV. RESULTATS

4.1 Exactitud global

Taula 5—*Exactitud global per model.*

Model	Exactitud	Encerts / 150
Qwen2.5-VL	70,7%	106
InternVL2.5	58,0%	87
Phi-3.5	45,3%	68

Qwen2.5-VL lidera amb un avantatge clar (+12,7 pp sobre InternVL i +25,4 pp sobre Phi-3.5). El seu 70,7% se situa per sobre del 62,99% de GPT-4V reportat a IllusionVQA, tot i que els benchmarks no són directament comparables (Fig. 1).

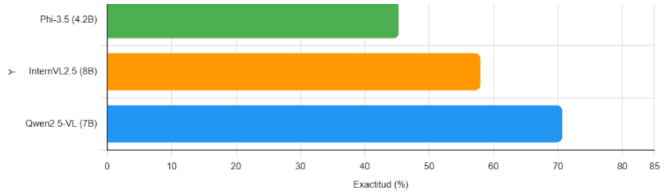


Fig. 1. Exactitud global per model.

4.2 Exactitud per categoria

Taula 6—*Exactitud per categoria i model.*

Categoria	InternVL	Phi-3.5	Qwen2.5-VL
Ambiguous depth	52%	60%	64%
Bulging	64%	36%	64%
Crater illusion	76%	72%	100%
Impossible figures	44%	32%	56%
Leaning Tower	80%	36%	92%
Reverspective	32%	36%	48%

Crater illusion és la categoria més fàcil: Qwen2.5-VL l'encerta perfectament (100%). Reverspective és la més difícil per a tots tres models (32–48%). Leaning Tower presenta una asimetria marcada: InternVL (80%) i Qwen (92%) hi van bé, però Phi-3.5 (36%) hi falla greument, cosa que suggereix que comparar imatges idèntiques requereix una capacitat que escala amb la mida del model (Fig. 2).

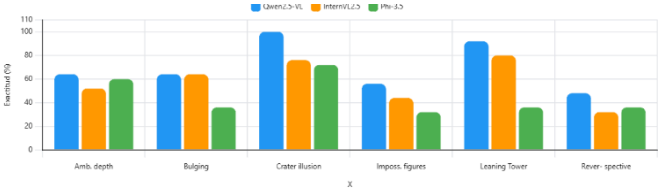


Fig. 2. Exactitud per categoria i model.

4.3 Taxa de lectura il·lusòria

La Taula 7 recull la taxa de lectura il·lusòria i literal de cada model.

Taula 7—*Taxa de lectura il·lusòria i literal global.*

Model	Taxa il·lusòria	Taxa literal
Qwen2.5-VL	28,7%	70,7%
InternVL2.5	35,3%	58,0%
Phi-3.5	38,0%	45,3%

Qwen2.5-VL és el que menys cau a la il·lusió (28,7%). Destaca que no tria mai l'opció il·lusòria a Crater illusion (taxa = 0%), cosa que confirma un domini complet d'aquesta categoria. En canvi, InternVL cau a la il·lusió el 68% de les vegades a Reverspective: «veu» la percepció humana però no la realitat física. Phi-3.5 presenta un equilibri quasi perfecte entre lectura il·lusòria (38%) i literal (45,3%), indici d'una comprensió feble (Fig. 3 i Fig. 4).

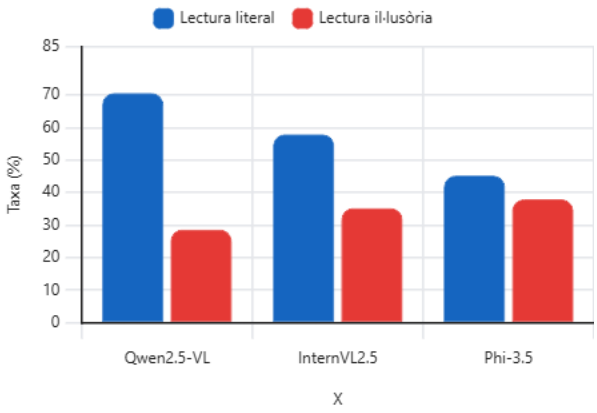


Fig. 3. Taxa il·lusòria vs. literal per model.

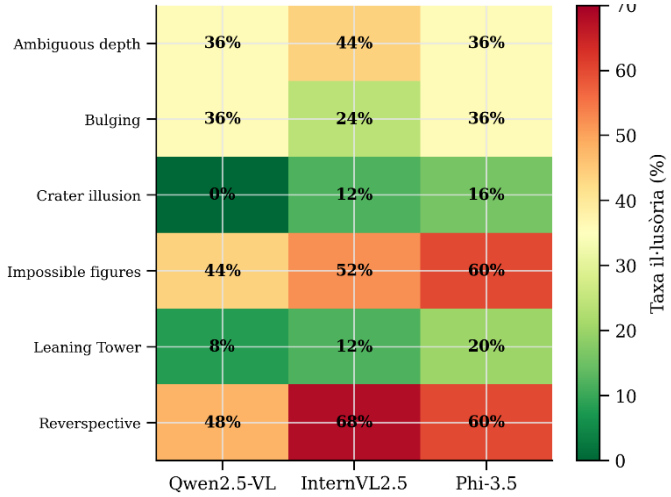


Fig. 4. Mapa de calor de la taxa il·lusòria per categoria i model.

4.4 Estabilitat: volatilitat del 100%

Taula 8—*Taxa de volatilitat per model.*

Model	Imatges volàtils / 30	Taxa de volatilitat
InternVL2.5	30/30	100%
Phi-3.5	30/30	100%
Qwen2.5-VL	30/30	100%

Cap model manté la mateixa resposta a les cinc variants de cap imatge. La fracció majoritària típica és de 0,4 (2 de 5 variants coincideixen) i l'entropia mitjana se situa entre 1,3 i 1,9 bits (sobre un màxim teòric de 2 bits amb quatre opcions), indici de respostes gairebé uniformement distribuïdes. Combinat amb el protocol blind, aquest resultat indica que l'encert aparent d'un model pot ser un artefacte de la interacció entre fraseig i posició de les opcions, i no una comprensió estable del contingut visual (Fig. 5).

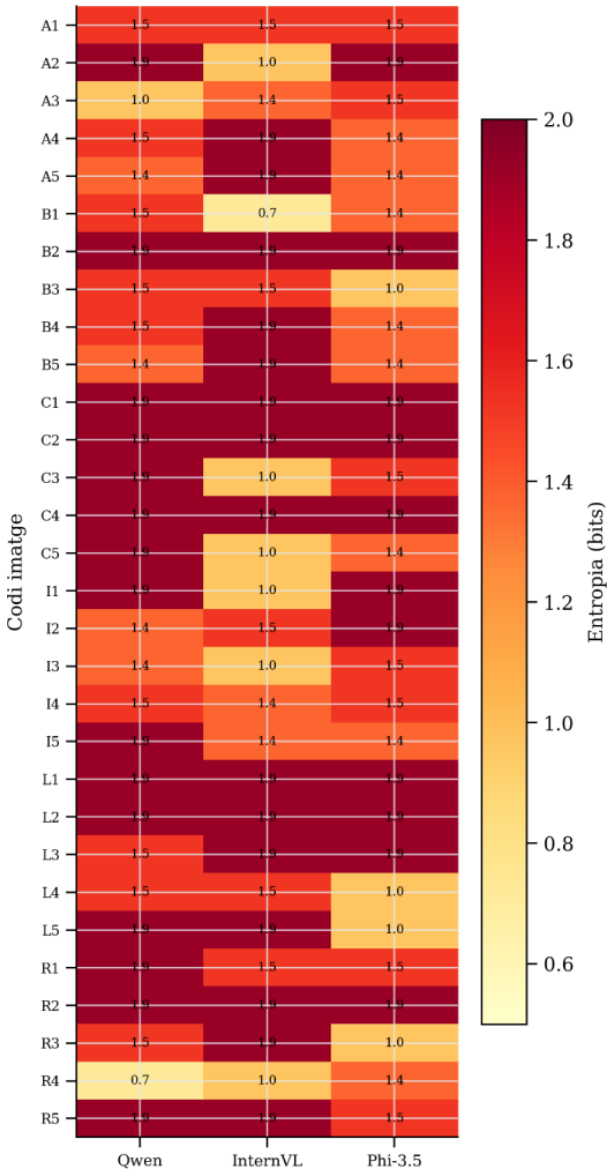


Fig. 5. Mapa de calor de l'entropia de Shannon per imatge i model (30 × 3).

4.5 Classificació dels errors

Taula 9—*Classificació dels errors per model.*

Model	Errors totals	Il·lusoris	Distractors	% il·lusori
Qwen2.5-VL	44	43	1	97,7%
InternVL2.5	63	53	10	84,1%
Phi-3.5	82	57	25	69,5%

Quan Qwen2.5-VL s'equivoca, gairebé sempre (97,7%) tria l'opció il·lusòria: comet «errors humans». Phi-3.5, en canvi, comet un 30,5% d'errors distractors, indici d'una comprensió visual més superficial. A Reverspective, InternVL comet el 100% d'errors il·lusoris (17/17): cada cop que falla, tria la lectura perceptiva humana. A Leaning Tower, Phi-3.5 presenta la taxa d'error il·lusori més baixa (31,25%), amb 11 dels 16 errors classificats com a distractors (Fig. 6 i Fig. 7).



Fig. 6. Classificació dels errors: il·lusoris vs. distractors.

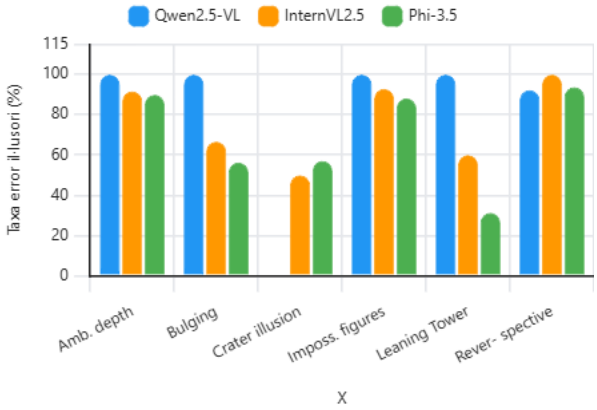


Fig. 7. Proporció d'errors il·lusoris per categoria i model.

4.6 Latència

Taula 10—*Latència per model.*

Model	Mitjana (s)	Mediana (s)	Total 150 (s)
Phi-3.5 (4,2B)	0,242	0,198	36,3
InternVL (8B)	0,374	0,241	56,1
Qwen2.5-VL (7B)	0,391	0,344	58,6

Phi-3.5 és el més ràpid, coherent amb la seva mida menor. Tots tres completen les 150 crides en menys d'un minut, cosa que confirma la viabilitat de l'entorn local per a experiments d'aquesta escala (Fig. 8).

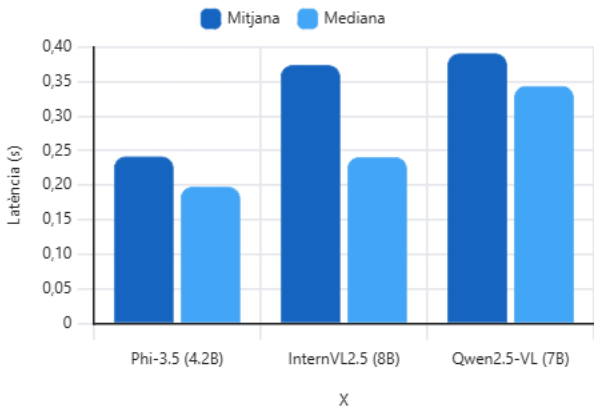


Fig. 8. Latència mitjana i mediana per model.

V. DISCUSSIÓ

5.1 Exactitud vs. comprensió robusta

El millor model (Qwen2.5-VL, 70,7%) supera el rendiment de GPT-4V a IllusionVQA (62,99%), però queda molt per sota del rendiment humà esperat (>90%). La volatilitat del 100%, però, qüestiona la significativitat de l'exactitud bruta: si el model canvia de resposta amb cada reformulació, el 70,7% reflecteix la probabilitat de topar amb un fraseig favorable, no una comprensió estable. Això planteja un problema metodològic de fons per a tots els benchmarks de VLMs: un sol valor d'exactitud pot ser profundament enganyós sense una mesura complementària d'estabilitat.

5.2 Sensibilitat al prompt i volatilitat de les respostes

La volatilitat és, possiblement, la contribució més rellevant d'aquest treball. Els estudis previs [1], [2] avaluen amb una sola formulació per imatge i assumeixen implícitament que el resultat és estable. El nostre disseny amb cinc variants i protocol blind demostra que aquesta assumpció no se sosté: l'encert d'un VLM davant d'una il·lusió depèn tant de la formulació lingüística com del contingut visual. L'observació connecta amb la troballa d'IllusionVQA segons la qual l'ICL i el CoT degraden Gemini-Pro: la sensibilitat al fraseig és un problema sistèmic dels VLMs davant d'estímuls perceptivament ambigus. El fet que tres models de famílies completament diferents (Alibaba, OpenGVLab, Microsoft) presentin la mateixa taxa de volatilitat (100%) apunta que aquesta inestabilitat afecta els VLMs amb independència de l'arquitectura concreta.

5.3 Gradient de dificultat per categories

L'anàlisi per categories revela un gradient coherent: (1) Crater illusion (72–100%) és la més fàcil —l'ombrejat és una pista de baix nivell ben captada—; (2) Leaning Tower (36–92%) requereix comparació espacial i escala amb la mida; (3) Bulging i Ambiguous depth (36–64%) tenen dificultat intermèdia; (4) Impossible figures (32–56%) demana raonament geomètric abstracte; i (5) Reverspective (32–48%) és la més difícil, perquè la inversió de profunditat supera tots els models. Sembla, doncs, que els VLMs processen millor les pistes perceptives de baix nivell (ombrejat, contrast) que les que exigeixen raonament espacial o geomètric d'ordre superior, en consonància amb la hipòtesi d'IllusionVQA.

5.4 Errors «humans» vs. errors arbitraris

La classificació dels errors aporta una dimensió qualitativa que transcendeix l'exactitud numèrica. Que Qwen2.5-VL cometi gairebé exclusivament errors il·lusoris (97,7%) indica que, quan falla, «percep» la il·lusió com un humà: les seves representacions visuals capturen propietats perceptives similars a les humanes, però no les superen. Phi-3.5, amb un 30,5% d'errors distractors, mostra una comprensió menys sofisticada: no només cau a la il·lusió, sinó que de vegades no entén l'escena en absolut. La distinció entre «error intel·ligent» (il·lusori) i «error per confusió» (distractor) podria ser una mètrica útil per a futurs benchmarks de percepció visual.

5.5 Relació entre mida del model i rendiment

Amb només tres models no es pot establir una relació estadística robusta, però s'observa una tendència: Qwen (7B) > InternVL (8B) > Phi (4,2B). El fet que InternVL (8B) no superi Qwen (7B) indica que l'arquitectura i l'entrenament importen tant o més que la mida bruta. La diferència és especialment marcada a les categories difícils

(Reverspective, Impossible figures), on la capacitat de raonament espacial sembla el factor discriminant. Pel que fa a la latència, Phi-3.5 (4,2B) és el més ràpid (0,242 s/crida), però la diferència absoluta és petita (< 0,15 s respecte dels altres), cosa que confirma que tots tres models són viables per a l'avaluació local.

VI. CONCLUSIONS

6.1 Conclusions principals

Tornant a la pregunta inicial —què veu realment un VLM quan la imatge ha estat construïda per enganyar la percepció—, les troballes d'aquest treball es resumeixen en cinc punts:

1) **Cap VLM comprèn les il·lusions de manera robusta.** El millor model (Qwen2.5-VL, 70,7%) queda molt per sota del rendiment humà (>90%) i la volatilitat del 100% indica que l'encert és inestable. Això confirma, en un entorn local i amb models de codi obert, la conclusió central d'IllusionVQA per a models propietaris.
2) **La reformulació del prompt altera sistemàticament la resposta.** Les cinc variants amb protocol blind demostren que els VLMs no posseeixen una representació estable de les il·lusions: la resposta depèn de la interacció entre fraseig i posició de les opcions. És la contribució principal del treball. Aquest fenomen coincideix amb la literatura prèvia sobre sensibilitat al prompt, que ja ha documentat variacions notables de rendiment davant reformulacions semànticament equivalents en altres tipus de tasques i dades [16]; la nostra aportació és evidenciar-lo específicament en la percepció d'il·lusions òptiques per VLMs.
3) **Els errors són majoritàriament «humans».** Quan els models fallen, tendeixen a triar l'opció il·lusòria (fins al 97,7% dels errors en Qwen2.5-VL), cosa que indica que les seves representacions visuals capturen propietats perceptives similars a les humanes, però no les superen.
4) **Hi ha un gradient de dificultat.** Les il·lusions basades en ombrejat (Crater) són les més fàcils; les que requereixen raonament geomètric (Impossible figures, Reverspective) són les més difícils.
5) **L'exactitud sola és insuficient per avaluar la percepció visual.** La combinació d'exactitud, estabilitat i tipologia d'errors ofereix una visió molt més rica i honesta del rendiment.

6.2 Línies de continuació

1) Ampliar el banc a més categories i més imatges per categoria, per permetre inferència estadística robusta amb intervals de confiança.
2) Separar els factors de sensibilitat al fraseig i a la posició de les opcions mitjançant un disseny factorial (fraseig fix × posició variable i viceversa).
3) Avaluar models més recents (Qwen3.5, GPT-4o, Gemini 2.0) quan l'entorn ho permeti, per determinar si les noves generacions superen la volatilitat observada.
4) Incorporar un baseline humà controlat amb els mateixos estímuls i format per comparar-hi directament.
5) Explorar l'explicabilitat: analitzar els mapes d'atenció dels models per entendre quines regions de la imatge consulten en cada categoria d'il·lusió.

VII. LIMITACIONS

1) **Models condicionats per l'entorn.** La restricció del driver CUDA 12.4 ha impedit usar models més recents i potents (Qwen3.5, GLM-4.6V); els resultats reflecteixen models de generació anterior.
2) **Mostra reduïda.** 30 imatges × 5 variants donen volum suficient per detectar patrons, però no per a inferència estadística robusta (intervals de confiança amplis).
3) **Absència de baseline humà propi.** No s'ha recollit un rendiment humà controlat; les comparacions es fan contra els resultats publicats d'IllusionVQA.
4) **Protocol blind i volatilitat.** La barreja de posicions, tot i ser un bon mecanisme anti-biaix, fa que la volatilitat observada pugui provenir tant de la sensibilitat al fraseig com de la sensibilitat a la posició; un disseny factorial futur podria separar tots dos factors.
5) **Idioma únic.** Tots els prompts s'han formulat en català. Atès que els VLMs són sensibles a la formulació lingüística, els resultats podrien variar en altres idiomes; caldria replicar l'experiment en anglès o en altres llengües per avaluar-ne la generalitat.

AGRAÏMENTS

BIBLIOGRAFIA

[1] A. Shahgir, K. Sayeed, R. Bhattacharjee et al., "IllusionVQA: A Challenging Optical Illusion Question-Answering Dataset for Vision-Language Models," Proc. COLM, 2024.

[2] J. Ngo, S. Sankaranarayanan, P. Isola, "Is CLIP Fooled by Optical Illusions?," Proc. ICLR (Tiny Papers), 2023.

[3] D. Geng, I. Park, A. Owens, "Visual Anagrams: Generating Multi-View Optical Illusions with Diffusion Models," arXiv:2311.17919, 2024.

[4] Y. Feng et al., "Illusion3D: 3D Multiview Illusion with 2D Diffusion Priors," arXiv:2403.11807, 2024.

[5] A. Kitaoka, "Akiyoshi's illusion pages," Ritsumeikan University. Disponible: http://www.ritsumei.ac.jp/~akitaoka/index-e.html

[6] Qwen Team, "Qwen2.5-VL: A Visual Language Model," Hugging Face, 2024.

[7] OpenGVLab, "InternVL2.5: An Open Multimodal Model," Hugging Face, 2024.

[8] Microsoft, "Phi-3.5-vision-instruct," Hugging Face, 2024.

[9] vLLM Project, "vLLM: Easy, fast, and cheap LLM serving," GitHub, 2024.

[10] Z. Zhao, E. Wallace, S. Feng, D. Klein, S. Singh, "Calibrate Before Use: Improving Few-Shot Performance of Language Models," Proc. ICML, 2021.

[11] OpenAI, "GPT-4V(ision) System Card," OpenAI, 2023.

[12] Gemini Team, Google, "Gemini: A Family of Highly Capable Multimodal Models," arXiv:2312.11805, 2023.

[13] H. Liu, C. Li, Q. Wu, Y. J. Lee, "Visual Instruction Tuning," Proc. NeurIPS, 2023.

[14] A. G. Shapiro, D. Todorović (Eds.), The Oxford Compendium of Visual Illusions. New York, NY: Oxford University Press, 2017.

[15] C. E. Shannon, "A Mathematical Theory of Communication," The Bell System Technical Journal, vol. 27, no. 3, pp. 379–423, 1948.

[16] M. Sclar, Y. Choi, Y. Tsvetkov, A. Suhr, "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design," Proc. ICLR, 2024.

Apèndix A — Criteri d'integració del material complementari

Aquest apartat justifica on s'ha situat cada tipus de material. La decisió respon a dos criteris: que el cos principal sigui llegible i que el treball continuï sent reproduïble.

Taula 11—*Criteri d'integració del material complementari dins la memòria.*

Material	Ubicació	Justificació
Memòria principal del TFG	Cos del document	Pregunta de recerca, estat de l'art, metodologia, resultats, discussió, conclusions i limitacions.
Documentació tècnica d'execució	Annex C	Material necessari per reproduir l'entorn, però massa operatiu per al cos principal.
Dataset complet amb solucions	Annex B i material adjunt	Fitxer de ground truth; al document s'expliquen les columnes i el criteri de construcció.
Banc blind de prompts	Annex B i material adjunt	Fitxer que reben els models, sense solucions, amb el protocol blind.
Scripts Python	Annex D i material adjunt	Base de la reproductibilitat; al document s'explica la funció de cada script.
CSV de resultats i mètriques	Annex E i material adjunt	Permeten auditar els càlculs; al cos només es mantenen les taules resum.

Apèndix B — Estructura del dataset i del banc blind

El disseny del dataset separa dues versions: una versió completa amb les respostes correctes i les etiquetes internes, i una versió blind que és la que reben els models durant l'experiment. Aquesta separació evita que el procés d'inferència tingui accés a la solució.

La versió completa conté 150 files i 17 columnes. La versió blind conté 150 files i 10 columnes. En tots dos casos, el volum correspon a 30 imatges multiplicades per 5 variants de prompt.

Taula 12—*Diccionari de columnes del dataset i del banc de prompts.*

Columna	Funció dins l'experiment
id	Identificador únic de cada pregunta, format per codi d'imatge i variant.
code	Identificador de la imatge, per exemple A1, B3 o R5.
category	Categoria perceptiva de la il·lusió.
variant	Variant del prompt: V0 base, V1 sinònims, V2 reformulació indirecta, V3 registre col·loquial i V4 marc instruccional.
image_file	Nom normalitzat de la imatge utilitzada a l'experiment, per exemple A1.jpg.
prompt	Pregunta formulada al model.
option_A/B/C/D	Opcions de resposta en format múltiple choice.
reading_A/B/C/D	Etiqueta interna de cada opció: literal, il·lusòria o distractora. Només apareix al dataset complet.
correct_option	Lletra de l'opció literal correcta. Només apareix al dataset complet.
illusory_option	Lletra de l'opció que representa la percepció il·lusòria humana. Només apareix al dataset complet.
ground_truth	Explicació breu de la realitat física de l'estímul. Només apareix al dataset complet.

La clau metodològica és que la resposta correcta i la resposta il·lusòria no ocupen sempre la mateixa lletra. Això obliga el model a llegir i interpretar les opcions, i redueix el risc que l'encert provingui d'un biaix posicional.

Apèndix C — Documentació tècnica d'execució

Aquest annex resumeix l'entorn final amb què s'ha executat l'experiment i les decisions que han condicionat la selecció dels models. La documentació completa d'execució queda com a material complementari.

Taula 13—*Entorn d'execució final de l'experiment.*

Component	Valor	Comentari
Màquina	VM VM_IA_03	Servidor amb GPU NVIDIA.
Driver GPU	CUDA 12.4	Límit que condiciona la pila de programari.
Gestor d'entorns	Miniconda / conda	Entorn aïllat per a l'experiment.
Entorn	vllm124	Cal activar-lo abans d'executar.
Python	3.12	Versió de l'entorn vllm124.
Motor d'inferència	vLLM 0.8.0	Servidor compatible amb l'API d'OpenAI al port 1234.

Component	Valor	Comentari
PyTorch	2.6.0+cu124	Compilat per a CUDA 12.4.
Transformers	4.49.0	Fixat per compatibilitat amb vLLM 0.8.0.
Dependències scripts	pandas, requests	Lectura de CSV, crides HTTP i anàlisi de resultats.

La regla fonamental és que la versió de CUDA del PyTorch ha de ser menor o igual que la del driver disponible. Per aquest motiu es va conservar vLLM 0.8.0 amb PyTorch cu124, en lloc d'actualitzar a versions més noves que exigien una CUDA superior.

Taula 14—*Models finals seleccionats.*

Àlies	Repo Hugging Face	Mida	Família
qwen25vl	Qwen/Qwen2.5-VL-7B-Instruct	7B	Alibaba
internvl	OpenGVLab/InternVL2_5-8B	8B	OpenGVLab
phi35v	microsoft/Phi-3.5-vision-instruct	4,2B	Microsoft

Els models inicialment previstos es van descartar per motius tècnics: Qwen3.5 i GLM-4.6V eren massa recents per a vLLM 0.8.0, mentre que Gemma 3n E4B no disposava de suport de visió en aquesta configuració. La decisió reforça la reproductibilitat perquè tots els models finals van poder executar-se realment amb el mateix protocol.

Apèndix D — Scripts i reproductibilitat

Taula 15—*Paper de cada script dins del flux experimental.*

Script	Funció
validate_dataset.py	Valida columnes, valors nuls, opcions vàlides, coherència entre opció correcta i lectura literal, distribució de variants i existència de les imatges.
run_experiment_vllm.py	Executa les 150 preguntes contra un model servit amb vLLM, envia la imatge en base64 i desa resposta, lletra, tokens, latència i errors.
analyze_results.py	Uneix resultats i ground truth, calcula exactitud, taxa il·lusòria, consistència, entropia, volatilitat, classificació d'errors i latència.
generate_figures.py	Genera els gràfics finals de la memòria a partir de les mètriques calculades.

La seqüència d'execució es basa en servir un únic model cada vegada, executar el banc complet, aturar el servidor i repetir el procés amb el model següent. Això evita conflictes de VRAM i garanteix que cada model rebi exactament el mateix conjunt de preguntes.

```
source ~/miniconda3/etc/profile.d/conda.sh
conda activate vllm124
cd ~/tfg_run
python3 validate_dataset.py --prompts Dataset_complert.csv --images_dir ./imatges
vllm serve Qwen/Qwen2.5-VL-7B-Instruct --served-model-name qwen25vl --port 1234 --max-model-len 8192 --trust-remote-code
python3 run_experiment_vllm.py --model qwen25vl --images_dir ./imatges --prompts banc_prompts_blind.csv --output
resultats/resultats_qwen25vl.csv --base_url http://localhost:1234
python3 analyze_results.py --results resultats/resultats_qwen25vl.csv resultats/resultats_internvl.csv
resultats/resultats_phi35v.csv --ground_truth Dataset_complert.csv --output_dir analisi/
```

Apèndix E — Fitxers de resultats i mètriques

El cos principal de la memòria presenta només els resultats agregats. Els fitxers següents permeten auditar el càlcul i reproduir les taules i figures.

Taula 16—*Fitxers de sortida i mètriques generades.*

Fitxer	Contingut
resultats_qwen25vl.csv	Respostes completes de Qwen2.5-VL per a les 150 crides.
resultats_internvl.csv	Respostes completes d'InternVL2.5 per a les 150 crides.
resultats_phi35v.csv	Respostes completes de Phi-3.5-vision per a les 150 crides.
resultats_complets.csv	Unió dels tres fitxers de resultats amb el ground truth i la lectura de cada resposta.
metrics_accuracy.csv	Exactitud global i per categoria.
metrics_illusory_rate.csv	Taxa de lectura il·lusòria i literal.

Fitxer	Contingut
metrics_consistency.csv	Fracció majoritària, entropia i volatilitat per imatge.
metrics_error_classification.csv	Classificació dels errors en il·lusoris, distractors i altres.
metrics_latency.csv	Mitjana, mediana, mínim, màxim i temps total de latència.
metrics_volatility.csv	Taxa de volatilitat global i per categoria.
report.txt	Resum textual de les mètriques principals.

Apèndix F — Resum auditable de resultats

Aquest resum reprodueix els valors centrals que sustenten les conclusions del treball:

- Exactitud global: Qwen2.5-VL = 70,7% (106/150), InternVL2.5 = 58,0% (87/150) i Phi-3.5 = 45,3% (68/150).
- Taxa de lectura il·lusòria global: Qwen2.5-VL = 28,7%, InternVL2.5 = 35,3% i Phi-3.5 = 38,0%.
- Volatilitat global: 100% en els tres models, és a dir, 30/30 imatges presenten canvis de resposta entre variants.
- Errors il·lusoris: Qwen2.5-VL = 97,7% dels errors, InternVL2.5 = 84,1% i Phi-3.5 = 69,5%.
- Latència total per 150 crides: Phi-3.5 = 36,3 s, InternVL2.5 = 56,1 s i Qwen2.5-VL = 58,6 s.

Apèndix G — Com s'haurien d'entregar els materials

Per a l'entrega final, la memòria en DOCX/PDF hauria d'anar acompanyada d'una carpeta de material complementari. Aquesta carpeta permet que el tribunal pugui verificar l'experiment sense que la memòria principal quedi saturada de codi o taules llargues.

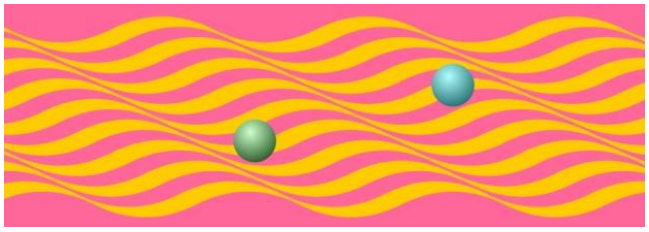
```
material_complementari/
  dataset/
    Dataset_complert.csv
    banc_prompts_blind.csv
  scripts/
    validate_dataset.py
    run_experiment_vllm.py
    analyze_results.py
    generate_figures.py
  resultats/
    resultats_qwen25vl.csv
    resultats_internvl.csv
    resultats_phi35v.csv
    resultats_complets.csv
    metrics_*.csv
    report.txt
  documentacio/
    Documentacio_TFG_Execucio.pdf
```

Aquesta organització separa clarament la memòria, que explica i interpreta, del material complementari, que permet reproduir i auditar.

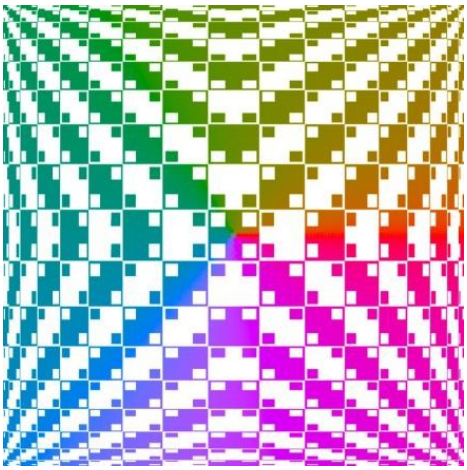
Apèndix H — Imatges utilitzades a l'experiment

Es mostra una imatge representativa de cadascuna de les sis categories d'il·lusió emprades a l'experiment. El conjunt complet de 30 estímuls (5 imatges per categoria) s'inclou al material complementari. Les imatges procedeixen principalment de la col·lecció d'Akiyoshi Kitaoka [5].

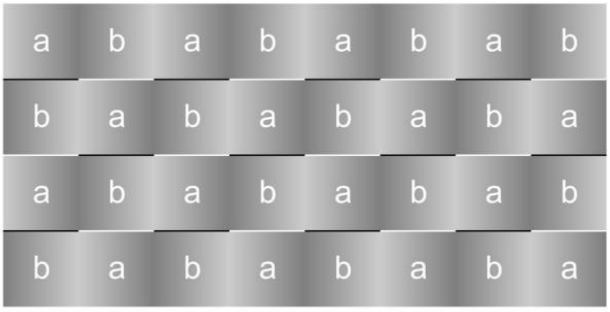
Taula 17—*Imatge representativa de cada categoria d'il·lusió.*



A — Ambiguous depth



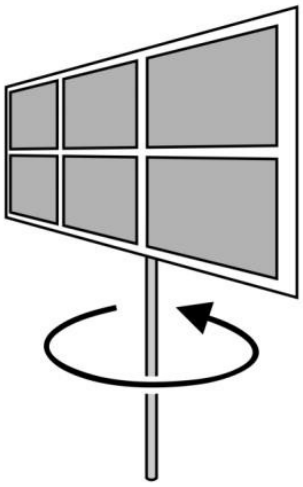
B — Bulging



C — Crater illusion



I — Impossible figures



R — Reverspective



L — Leaning Tower